UNIVERSITÀ DEGLI STUDI
DI TRENTO

DEPARTMENT OF INFORMATION ENGINEERING AND COMPUTER SCIENCE
ICT International Doctoral School

Detecting Brain Effective Connectivity with Supervised and Bayesian Methods

PhD candidate:
Danilo Benozzo

Advisor: Dr. Paolo Avesani[1,2]

Co-Advisor: Dr. Emanuele Olivetti[1,2]

[1]NeuroInformatics Laboratory (NILab), Bruno Kessler Foundation, Trento, Italy

[2]Center for Mind and Brain Sciences (CIMeC), University of Trento, Italy

April 2017

My own experience is that, unlike art,
causality is a concept whose definition
people know what they do not like but
few know what they do like.

GRANGER, C. W. J. (1980)

Abstract

The study of causality has drawn the attention of researchers from many different fields for centuries. In particular, nowadays causal inference is a central question in neuroscience and an entire body of research, called brain effective connectivity, is devoted to detecting causal interactions between distinct brain areas. Brain effective connectivity is typically studied by the statistical analysis of direct measurements of the neural activity.

The main purpose of this work is on methods for studying time series causality. More in details, we focus on a well-establish criterion of causality: the Granger criterion, which is based on the concepts of temporal precedence and predictability.

Firstly, we consider the standard parametric implementation of the Granger criterion that is based on the multivariate autoregressive model, where we face the problem of model identification. For this purpose, we present a new Bayesian method for linear model identification and we explore its capability of modeling the sparsity structure of the signals.

As a second contribution, we look at the causal inference through the lens of machine learning and we propose an approach based on the concept of learning from examples. Thus, given a set of signals, their causal interactions are estimated by a classifier that is trained on a synthetic dataset generated by a parametric model. This approach, that we call supervised parametric approach, is implemented by adopting the Granger criterion of causality and compared with the standard parametric measure of Granger causality. Moreover, the roles of the feature space and the generative model of the training set are investigated through a simulation study. Additionally, we show an example of application on rat neural recordings.

Finally, we focus on the bias introduced by parametric methods when applied in a real context, i.e. the inability of having a fully realistic generative model. For this purpose, we analyze how the supervised parametric approach can help in making the inference more application-dependent, by exploiting a physiologically plausible generative model.

Keywords

Causal Inference, Time Series Causality, Granger Causality, Effective Connectivity, Causal Connectivity, Bayesian Linear Regression Method.

Conference papers

- E. Olivetti, D. Benozzo, S.M. Kia, M. Ellero and T. Hartmann, *The Kernel Two-Sample Test vs. Brain Decoding*, Proceedings of the 2013 International Workshop on Pattern Recognition in Neuroimaging (June 2013), pp. 128-131
- D. Benozzo, E. Olivetti, P. Avesani, *Classification-Based Causality Detection in Time Series*, In Machine Learning and Interpretation in Neuroimaging, Vol. 9444 (2016), pp. 85-93

Manuscripts under review

- D. Benozzo, P. Jylänki, E. Olivetti, P. Avesani, M. A. J. van Gerven, *Bayesian Estimation of Directed Functional Coupling from Neural Time Series*
- D. Benozzo, E. Olivetti, P. Avesani, *Supervised Causal Graph Estimation*

Manuscript in preparation

- D. Benozzo, J. Bin, S. Panzeri, P. Avesani, *Validating Unsupervised and Supervised Brain Connectivity Inference Methods with Realistic Neural Network Simulations*

Software

- GMEP: Bayesian method for linear model identification and brain causal connectivity analysis
  `https://github.com/ccnlab/GMEP`
- Supervised approach for causal graph estimation
  `https://github.com/danilobenozzo/supervised_causality_detection`

Contents

List of Tables

List of Figures

Part I

Dissertation

Chapter 1

Introduction

In our daily life, the observation of events largely contributes to build our experience and in particular, our attention is drawn to events that show some sort of *relation*. Like a fire with the temperature of the air around, the tides of the sea with the Moon's phases, the weather conditions and the atmospheric pressure etc. In all these events, the variables are *dependent* on each other.

Dependence is an essential component of statistic, pattern recognition and machine learning, and finding it between events is the purpose of a large body of science. We could model the dependence between events, make predictions on the behavior of certain variables of the system, interact with the system in order to reach a certain state. However, this does not explain how the dependence arises. The origin and the motivation of a dependence lie in the field of *causality*.

The study of causality is quite controversial and it involves different fields. Firstly, from philosophy, there are many contributions since the definition itself of causal interaction is not straightforward. Then from the side of the numerical sciences like mathematics, physics, computer science, engineering, the effort is in developing *methods* for the inference of causal relations between events. And this area is where this work takes place. Specifically, we will approach the problem of causal inference between *time series* with direct attention to *neuroscience* applications.

Nowadays, acquiring multivariate *neural recordings* is a common practice in neuroscience experiments, thanks to the technological improvements in measuring devices. The *multivariate* nature of the recordings and the fact that the acquisition is concurrent between different brain regions have shifted the attention towards *brain connectivity*. And this raises the need for adequate analysis methods.

Brain connectivity aims to study the pattern of interactions exhibited among distinct

brain units within the brain [27]. This type of study can be conducted at different levels of scale and according to the adopted scale, the concept of brain unit changes. For example, brain connectivity can be studied from the microscopic level of single synaptic connections to the macroscopic level of brain regions. Moreover, depending on the type of interactions of interest, brain connectivity is divided into *structural*, *functional* and *effective* connectivity. In the first case, the connectivity patterns are referred to the anatomical links i.e. the neural pathways, in the second case to the statistical dependencies between brain activity in different units and in the last one to the causal interactions between them [44]. In particular, effective connectivity provides information about the direct influence that one or more units exert over another and aims to establish causal interactions among them [17].

We will see that, in order to derive conclusions about the effective connectivity, there is a necessary condition to fulfill. And it regards the capability of modeling the system considering both the physiological structure and also its dynamics. This constraint considerably increases the complexity of the problem, since it requires a deep knowledge of the system and an explicit definition of the causal effect.

So, on the side of our application scenario, there is the need of a parametric approach able to accurately model the system in order to facilitate the interpretation of the inference. While on the side of methods, an entire area of research has been developed to measure causality by starting from the concepts of *causal calculus* and *interventions* [36]. This type of analysis assumes the possibility of manipulating the system, thus it is not a standard statistical analysis.

The direct application of the interventionist approach in neuroscience is not straightforward. For example, it is prevented by ethical limitation on perturbing the system and by the experimental paradigm commonly adopted in neuroscience. Moreover, it does not allow the investigation of the inner working of the system [1] as effective connectivity requires.

The common way to study direct interactions from brain recording is through the so-called *causal connectivity analysis*. This lies in the area of statistical analysis since it assesses parameters of a distribution from samples that are supposed to be drawn of that distribution [37]. Thus, this is not a causal analysis but to be precise we should refer to it as an estimate of a weaker form of causality or of a directed functional coupling (in neuroscience terms). Probably the most used criterion of (weak) causality in causal connectivity analysis is the *Granger criterion* [22]. The Granger criterion is based on the assumption of precedence and predictability of the cause with respect to its effect.

Precedence means that a cause has to temporally precede its effect. Predictability is referred to the conditional dependence that exists between the past of the causes and the future of the effect, conditioned on the past of the effect itself.

The focus of this thesis is on methods for causal connectivity analysis. Thus, given a set of signals simultaneously recorded from distinct brain units, we seek for causal connections among them, i.e. the causal structure. We consider as causal connection a directed binary relation that carries only qualitative information [14], in the sense that we do not quantify the strength of the interaction.

1.1 Problem statements

The standard way in which a causal connectivity method is developed starts with the implementation of a criterion of causality. The implementation may assume a specific stochastic model for the underlying process of data generation, in this case, we refer to a parametric formulation of the criterion. Or in the case of a model-free approach, the formulation is said to be non-parametric. We label these two approaches as unsupervised.

Focusing on the parametric implementation of the Granger criterion, the model usually adopted is the multivariate autoregressive (MAR) model. And in order to derive the corresponding causal measure, such model has to be identified from the observed data.

As we will see in Section 3.1, the MAR model can be identified by solving a multivariate linear regression problem. A large number of solutions are available in the literature for this type of problem but here we will focus on two relevant aspects of our domain of application. Firstly, due to the nature of neuroscientific dataset, the number of unknowns can be massive thus the need of *regularizing* the inference to overcome issues like over-fitting and non-uniqueness of the solution. Secondly, there may be *prior knowledge* on the causal structure that we would like to include in the inference process. To this purposes, we seek for a solution that is *Bayesian-based* and that regularizes the inference by a *structured prior*, i.e. by enforcing the sparsity structure of the unknown coefficients.

Having a more robust method for linear identification is of interest in the context of the parametric Granger causal analysis but it still requires to adopt the MAR model in the implementation of the criterion. On one hand, this assumption simplify the issue of the model identification since in principle it is an easy task and well studied in the literature. On the other, in some cases the MAR model suffers for lack of realism because it does not consider the actual mechanism of signal generation. A straightforward alternative could be the adoption of a more realistic model but we may face difficulties in inverting it. To

this purpose, we propose a different approach compared with the unsupervised ones and it is based on the concept of learning the causal structure from examples. In other words, the proposed approach lies in the area of supervised learning, indeed we will refer to it as the *parametric supervised* approach. Thus, i) we will define a *feature space* in which each set of measurements is mapped, ii) a *generative model* will be adopted for the training dataset and iii) we will design the *classification schema* to infer the connections of the causal structure given a set of time series.

The supervised approach will be initially tested by customizing it to the case of a MAR model and designing the features space on the Granger criterion of causality. This may seem a contradiction since we motivated the supervised approach by stressing the possibility to use a more realistic model for describing the stochastic process. But we firstly want to study the idea itself of inferring the causal structure through a learning phase by focusing only on the model identification problem and on the feature space, without exploiting more accurate models. Moreover, this choice allows a comparison with the standard unsupervised approach since both are built on the same assumptions.

Afterworld, we will present a dedicated activity that aims to study the case in which a more plausible model is used instead of the MAR model. We introduce a neuro-physiological model for the stochastic process and we used it both to evaluate the proposed approach for causal inference and also as generative model for the training dataset. About the feature space, it will be still kept as before, so based on the Granger criterion. This is possible because of two reasons: i) in the supervised approach the training phase has its own stochastic process which does not directly depend on the causal criterion that instead is implemented in the feature space and ii) thanks to the training phase, the model identification is not required thus the complexity of the generative model is not a potential problem.

1.2 Contributions

The thesis is divided into three parts, that represent the three main activities of my PhD research. These three parts refer to the problems presented in Section 1.1. Here, we will recall them and give a summary of the proposed solutions.

In the first part, we propose a new Bayesian method for linear model identification with a *structured* prior (GMEP). Our aim is to apply it as linear regression method in the context of the parametric Granger causal inference. GMEP assumes a Gaussian scale mixture (GM) distribution for the group sparsity prior and expectation propagation (EP)

is used for approximating the posterior inference. The proposed method is investigated both on simulated and empirical data, from which the advantages given by the flexibility in defining the sparsity structured prior, become evident.

The second part introduces a new perspective to look into the problem of causal inference, which we call *supervised parametric* approach. The supervised approach is based on machine learning and, specifically on learning from examples. As example, we refer to a set of time series together with their true causal structure. A classifier is trained to identify causal interactions on a dataset generated by the chosen *generative model*. The dataset is mapped into the proposed *feature space*, that is based on the adopted criterion of causality. This approach is tested in the context of the Granger criterion of causality and in particular by adopting its multivariate autoregressive implementation. Multiple experiments are presented on simulated datasets, in order to investigate the properties of the proposed feature space and the role of the chosen generative model. An example of real application is shown on rat neural recordings.

In the last part, we focus on the importance of the generative model within the proposed supervised causality method. In particular, from the point of view of the neuroscientific interpretation, a causality measure can be interpreted in terms of effective connectivity only in the case of a stochastic process that physiologically models both the *structure* and the *dynamics* of the neural activity. Thus, this third activity is meant to investigate the effect of adopting a more physiologically plausible generative model for evaluating the proposed methods. In other words, the effect of violating the assumption of data generation assumed by the parametric implementation of the causal criterion. Moreover, this new model is also used for generating the training dataset for the learning phase of the supervised approach. We will show that, thanks to the supervised approach, we are in the position to accommodate a Granger-based parametric implementation of causality to a new generative model. In particular, we evaluated the case of the Granger criterion implemented in the feature space when applied on a generative neuronal model able to simulate the activity of a real neuronal network.

In conclusion, the main contributions of this thesis are:

- the characterization of GMEP as linear regression method with a group sparsity prior and its evaluation in the context of the causal inference;
- the development of a supervised method for causal graph estimation, its analysis under the Granger definition of causality and its comparison with the standard Granger index;
- the evaluation of the effect of violating the assumption made by the parametric

implementation of the Granger criterion, on the generative model and the analysis of how the supervised approach could reduce this effect.

1.3 Structure of the thesis

The thesis is divided into two main parts: Dissertation and Papers. The former part summarizes the three main contributions of my PhD research and it is structured as follows:

- Chapter 2 starts with an overview about causality focusing on the difficulty of formulating its definition and difference with correlation, then the temporal dimension is introduced and so the idea of time series causality. The Granger criterion is presented together with a review of the most common causal measures that have been derived from it. Finally, the chapter concludes with a series of background knowledge on the area of machine learning and neuroscience that will be used later.
- In Chapter 3 the problems of which we aim to provide a solution are stated. The chapter is structured in three sections one for each activity that we will present. The same structure is repeated also in the following chapters.
- Chapter 4 is about solutions. It focuses on the problems described above and starting from how they are commonly faced in the literature, our proposed solutions are described.
- Chapter 5 collects the outcomes of the analyses done to characterize and evaluate our proposed solutions. A series of experiments is presented together with their related results.
- Chapter 6 concludes the first part of the thesis. It recalls the main results presented before and comments them by focusing on the related implications, limitations and future works.

The second part contains the manuscripts related to the activities presented in the first part.

- Chapter 7 contains the manuscript titled *Bayesian Estimation of Directed Functional Coupling from Brain Recordings*, this work was conducted in collaboration with the *Computational Cognitive Neuroscience Lab* at the *Donders Institute for Brain, Cognition and Behaviour* (Nijmegen, NL) where I did my internship.

- In Chapter 8 there is the manuscript titled *Supervised Casual Graph Estimation*, this work is the results of an internal activity of the *NeuroInformatics Lab* that was initially born as an attempt to compete in the Causal2014 causal inference competition.
- To conclude, Chapter 9 has inside the manuscript *Validating Unsupervised and Supervised Brain Connectivity Inference Methods with Realistic Neural Network Simulations* that refers to a follow-up activity of the work in Chapter 8. Such manuscript describes a collaborative work with the *Neural Computation Lab* at the *Italian Institute of Technology*.

Chapter 2

Background

This chapter introduces the concept of causal inference starting from the problem of establishing a definition of causality and focusing on its strong link with correlation. Then, the time dimension is considered and so the problem of time series causality. After that, a section is dedicated to the Granger definition of causality and to methods that have been derived from its implementation. To conclude, since this thesis is a multidisciplinary work in which methods belonging to computer science and statistics are designed to be applied on neuroscience data, a short overview on machine learning and brain connectivity is given. In particular, we will focus on two central problems in the field of supervised learning that are regression and classification. Regarding the part dedicated to neuroscience, the main goal is to give a general introduction to brain connectivity with particular attention on the effective and causal connectivity.

2.1 Causality

The study of causality has drawn the attention of researchers for centuries from many different fields. In this section, we provide a general overview of some difficulties on defining the meaning of a causal link between events and we warn against naive conclusions that may be drawn from a correlation analysis.

2.1.1 Definition of cause of Aristotle and Hume

Defining what a cause is has been an interesting question since centuries. Aristotle gave one of the earliest definitions of cause by relating the concept of cause to a *why* question. So the meaning of cause was connected to an explanation on what the origin of the

phenomenon is, on what it is made from, or why it is done. In other words, identifying causes means understanding why one thing happened instead of another.

Among the many contributions that have been had after Aristotle, David Hume in the 18th century reshaped the problem by distinguishing between the *meaning* of cause and the *approach* to seek it. He gave to the process of finding causal relation a proper identity and a primary role in defining the meaning itself of causal relation. Indeed, according to Hume a causal relation is the relation that results as output of the process of causal inference. In particular, that process regards our perception of causal relation (causal sense) that is modulated by the observation of regular patterns of occurrences.

Thus in Hume's view, the inference process comes from experience and it is characterized by the temporal precedence of the cause and the contiguity between cause and effect both in time and space.

It is easy to see that Hume's work on causality is not true in general and many counterexamples show that it does not cover all possible cases. For instance, it does not recognize a causal interaction in situation in which the lack of a factor causes an effect, e.g. lack of vitamin C and scurvy.

2.1.2 Relation and differences between correlation and causality

Having related a causal interaction with the presence of regular patterns of occurrences may erroneously implies a strong link with *correlation*. Actually, causation is something more than correlation and deriving causal relations from correlation studies is not straightforward.

A first difference is that correlation is symmetric while causality might be not. Thus, causality has a *direction* and it emerges from psychological experiments that time together with prior knowledge plays a key role in the human perception of a causal direction.

Despite their difference, correlation is still commonly used to find relations between variables, the key point is in the experimental setup and on how correlation is computed and interpreted. This means that a large amount of work is on discarding correlations that emerge from events that are simply observed at the same time from correlations that really point out causality. As we stress that not always correlation implies causality, it is important to mention that there are cases of causal relations without correlation. Thus, correlation is neither a sufficient nor necessary condition for causality.

Some famous examples were formulated of correlated events in which the causal link is definitely suspicious: in [33] authors reported a surprisingly very high correlation between chocolate consumption per capita and Nobel prize assigned in a country, sleeping with

intense ambient light was linked to the development of myopia in children in [41], or the relation between the price of British bread and the level of Venetian seas as shown in [53]. The presence of a *confounder* or an *unmeasured common cause* may explain the correlation.

Differently, there might be causal relationship without an apparent correlation, e.g. running and weight since running might influence the appetite of the athlete then the correlation between running and weight depends on the strength of the interaction between these three variables.

2.1.3 Simpson's paradox

The definition of the set of *variables* of which we aim to infer the casual network deserves particular attention, since it may affect the final result. We will come back on this issue in Section 2.2, because it is of relevance for our analysis. At this stage, in which we generally refer to the problem of data causality, it is worthwhile to mention the so-called Simpson's paradox for having an idea on the importance of choosing the proper *partition* of variables in our data. Simpson's paradox refers to a phenomenon in which the association between a pair of variables changes or reverses sign if conditioned on a third variable, regardless of the value taken by this latter. A common example used to explain it, regards the analysis of the effect of two medical treatments on a population of patients. Looking at the entire population we may conclude that e.g. treatment A gives the highest recovery rate, while conditioning the analysis on a third variable, e.g. the gender of the patient, the conclusion may reverse and be treatment B the most effective for both males and females.

Our main aim in showing this effect is to underline the relevance of the partitioning of the data on the final results. In addition, the link of this effect with the field of causality is even stronger. Indeed in [38] an interesting analysis is done on this effect by proposing a characterization of the phenomenon that includes its explanation and resolution, in the context of the causal framework developed in [36].

Regarding a real application scenario, we commonly analyze data that are measured from a predefined *parcellation* of the system, e.g. the brain in our application. And such parcellation can be given by intrinsic constraint of the acquisition device, e.g. its spatial resolution or by anatomical a priori information. Being aware of this effect, the outcome of a causal analysis may be better interpreted.

2.2 Causality over time

Beyond the association that may emerge from the data, other higher-level relations should emerge in order to establish a causal order in the events. Here, we focus our attention on the dynamics of the events. Thus, we consider the *temporal dimension* to derive information on the causal order.

2.2.1 Time series causality

In the previous section, the problem of establishing a causal connection was posed among data that contains discrete events, e.g. consumption of chocolate versus number of Nobel prizes both evaluated across countries. This type of studies can be defined with the general term of data causality, whereas by restricting the field of application to the case of time series we refer to *time series causality*. Thus the dimension along with the causal relationship is evaluated is time. A relevant difference in time series causality analyses respect to the more general data causality analyses, is on the difficulty of performing *interventional* experiments. This motivates the debate in the literature about whether considering or not time series causal inference an actual causal analysis rather than a simpler statistical analysis [32]. The need to approach the problem through an *observational* framework is due to the complexity of the system under analysis. Indeed, this kind of analysis is normally applied in fields like neuroscience, astrophysics, climate changes etc. in which the direct intervention into the system dynamic is prevented by the complexity of the system or by ethical limitations on perturbing it.

The need of approaching the analysis through an observational way implies the assumption of complete *observability* of the system. Practically, when the causal interaction between two processes is evaluated we assume there is no hidden process which is a common driver of the two ones under analysis [14]. This assumption still holds for methods in which a forward model of the system is used since such model accounts for hidden states that are mapped to observed quantities but it does not treat the case of a hidden process that plays as a common driver. Moreover, another important aspect to consider is the stationarity of the processes since this is a common assumption for the applicability of many of the existing methods.

2.2.2 Time order and causal order

Having focused our attention on the temporal dimension, it is worthwhile spending few words on the relation between *time order* and *causal order*. First of all, it is a very

controversial problem from different perspectives, i.e. physics, philosophy, logics and a comprehensive overview is far beyond the scope of this thesis. The idea that time and causality are connected is something that we naturally perceive, and it was even used by Leibniz who firstly proposed to reduce time order to causal order.

The order of events influences how we link and explain facts and it is the base of how we infer causality from observations. It directly emerges from Hume's theory that the asymmetry of causal relation due to the constraint on the time order corresponds in pointing the arrow of causality on the same direction of the temporal arrow. This seems to lead to a comprehensive theory in which causal asymmetry and temporal asymmetric are the two sides of the same coin. But problems like simultaneous causality or retro-causality would be still open. From the physics point of view, causal order may be explained in thermodynamic terms: the convention of defining positive time (and so causal order) through growing entropy. This leads to the fact that causes (and not ends) determine the occurrences of the present. But among physicists, it is commonly accepted that laws that define universe are uniquely defined. Thus causality does not define a direction of time [26]. Despite this, the causal asymmetry is used everyday in science for making experiments and the reason why our causal sense is aligned with the temporal order is still not formalized.

2.2.3 The principle of common causes

The problem of determining the causal order has led the so-called Principle of Common Cause. This principle was a fundamental building block of the concept of probabilistic causality developed by Reichenbach in [52]. Intuitively, the basic idea of probabilistic causality is that a cause makes its effects more likely. More in details, the principle of Common Cause states that if two random variables are dependent then one of the following explanations holds: i) the two variables are causally connected, ii) there exists a third random variable which is a common cause of both, or iii) there exists a third random variable which is a common effect of both, upon which the observations are conditioned. Common causes are related to the presence of confounding variables. Moreover, a confounding variable which is also unobserved is called unobserved confounder and it may lead to *spurious* causal connections.

This approach of causality is not free from criticism. The fact that a dependence between random variables is explained in terms of causality may lead to wrong conclusions. Logical fallacies may derive from this approach like the so-called *post hoc ergo propter hoc* and *cum hoc ergo propter hoc*. The former refers to the risk of causally relating events

only on the basis of their time order. Two following events not necessarily are connected. The latter refers to the phenomenon of chance coincidence, i.e. two events that occur together do not necessarily have a common cause.

2.2.4 The problem of strong correlation

We consider here an additional issue related to the time series causality that is the high correlation exhibited by time series both in time and space. In studying complex system through measured multivariate time series, their auto- and cross-correlations may represent a starting point for evaluating the causal network. In particular, an entire group of tools for doing causal inference, called correlation causality tools, is based on the evaluation of the lagged cross-correlation. But as we already mentioned before, correlation can be due to many reasons. So it is crucial to disentangle between correlation that is associated to an actual causal interaction than spurious correlation. Spurious correlations may derive from a large number of sources, especially in the case of complex system.

A high correlation in time related to the actual dynamics of the system, may hide causal relations across signals. Moreover, a strong autocorrelation in time violates the assumption of independent samples commonly assumed by significance tests [42, 8].

Considering that our application scenario is neuroscience, it is worthwhile to underline the so-called effect of *volume conduction*. Volume conduction refers to the effects of recording electrical potentials at a distance from their source generator [43]. Regarding brain recording, the recording electrodes are not in direct contact with the neurons, except for single cell recordings. This implies that the activity of one source is recorded from multiple sensors. And this will affect the actual cross correlation of the sensors. In order to contrast this distortion, methods for projecting the sensor acquisition to the source space have been proposed [12].

2.3 Granger causality

This section is meant to introduce the *Granger criterion* of causality and its related *parametric* and *non-parametric* implementations. Moreover, regarding the parametric implementation the described measures are grouped according to their domain of definition, i.e. *temporal* and *spectral* domain.

2.3.1 Causal inference

Granger causality is one of the most widespread criteria for causal inference among brain recordings [50] and it is based on the assumptions of precedence and predictability of the cause with respect to its effect. As precedence we mean that a cause has to precede its effect and predictability is referred to the conditional dependence that exists between the past of the causes and the future of the effect conditioned on the own past of the effect. In the bivariate case, the criterion was originally enunciated as a condition of non-causality. Assuming that X and Y are two processes of which we record two time series $\{X\} = \{X_1, X_2, \ldots, X_N\}$ and $\{Y\} = \{Y_1, Y_2, \ldots, Y_N\}$ both of N time points, the criterion says that there is no causality from Y to X if

$$p(X_{t+1}|X^t) = p(X_{t+1}|X^t, Y^t), \quad \forall X^t, Y^t \tag{2.1}$$

where X^t and Y^t mean the past of the process up to time point t. We notice that the condition of non-causality is based on the comparison between probability distributions thus at this stage there is no constraint on the stationarity of the processes. Moving forward from the bivariate case and considering the potential interaction of a third process Z the criterion becomes

$$p(X_{t+1}|X^t, Z^t) = p(X_{t+1}|X^t, Y^t, Z^t), \quad \forall X^t, Y^t, Z^t \tag{2.2}$$

Generalizing, Z^t refers to the past of any other process that may interact with X and Y. The accuracy of a test derived from the Granger criterion is strongly dependent on the processes on which it is conditioned to. This is related to the assumption of complete observability as was mentioned in 2.2.1 due to the observational nature of this type of analysis.

A criterion to be applied has to be implemented. A criterion of causality defines which condition has to be satisfied to establish that two (or more) processes are (or are not) causally interacting. Given a certain criterion and according to how it is formulated, different measure of causality can be developed. There are cases in which the measure is defined by assuming a model for the underlying process of data generation, the so-called parametric formulations of the criterion. Or in the case of a model-free approach, the formulation is said to be non-parametric. Examples of implementations are provided below for both approaches.

2.3.2 Parametric implementations

Generally speaking, in the parametric formulation this pipeline is followed: a criterion of causality is chosen, according to it a model of the generative process is assumed and then a measure for causality is defined by considering the modelling assumptions. Commonly the computation of the causality measure requires the identification of the model and this step may be not trivial [58]. Moreover, to obtain the causal graph from the computed measures, the significance of the non-zero values needs to be tested. This can be done for example by bootstrapping techniques or by knowing the actual distribution under the null hypothesis. Regarding the Granger criterion, its standard parametric implementation assumes a linear multivariate autoregressive (MAR) modelling of the process. This assumption refers to how time series are interacting with each other, but without explicitly modelling the physical mechanism of generation. The autoregressive representation has led to different formulations of measures of causal interaction both in time and spectral domain.

Time domain We refer to the temporal formulation of the autoregressive implementation of the Granger criterion as the Geweke measure in time. Consider a system of three stationary stochastic processes X, Y and Z. The pair-wise conditional approach examines whether Y has a direct influence on X given the presence of Z by decomposing

$$X_t = \sum_{i=1}^{\infty} a_{xx,i} X_{t-i} + \sum_{i=1}^{\infty} a_{xy,i} Y_{t-i} + \sum_{i=1}^{\infty} a_{xz,i} Z_{t-i} + \varepsilon_{x,t} \tag{2.3}$$

Afterwards, the reduced autoregressive representation of X is considered

$$X_t = \sum_{i=1}^{\infty} a'_{xx,i} X_{t-i} + \sum_{i=1}^{\infty} a'_{xz,i} Z_{t-i} + \varepsilon'_{x,t} \tag{2.4}$$

The Geweke index of causality in time domain $F_{Y \to X|Z}$ evaluates which of the two regressions (2.3) and (2.4) models better the process X by computing

$$F_{Y \to X|Z} = \ln \frac{\Sigma'_{xx}}{\Sigma_{xx}} \tag{2.5}$$

where $\Sigma'_{xx} = \mathrm{var}(\varepsilon'_{x,t})$ and $\Sigma_{xx} = \mathrm{var}(\varepsilon_{x,t})$ are the residual variances of the MAR models (2.3) and (2.4) respectively. Equation (2.5) is interpreted as the variation in prediction error when the past of Y is included in the regression. A meaningful reduction of the residual variance when the candidate cause is included in the model identification, implies a better model for the effect. Thus, the time series evaluated as possible cause is said to Granger cause the time series evaluated as effect [9]. An important aspect is the

statistical significance of the estimated causal measure and the common practice is to look at Equation (2.5) as the test statistic of a log-likelihood ratio test. In particular, it results that under the null hypothesis of zero causality $H_0 : a_{xy,i} = 0, \forall i$ the Geweke measure has an asymptotic χ^2 distribution up to a scaling factor which depends on the sample size and with degree of freedom equals to the difference in the number of parameters between the models in Equations (2.3) and (2.4). Under the alternative hypothesis, the scaled test statistic has an asymptotic noncentral χ^2 distribution with noncentrality parameter that corresponds to the scaled casual measure. In a more general formulation the three processes may be multivariate thus they may represent a set of variables. It has been proved that this measure of causality since is based on the linear assumption of the process, is a test of Granger causality on the first moment statistic of the underlying probability distributions [23].

Spectral domain The autoregressive parametric formulation of the Granger criterion was also implemented in the spectral domain. It was introduced in [19] and named Geweke spectral measure of Granger causality. In the bivariate case, the Geweke spectral measure from X and Y at the frequency ω, is defined as the natural logarithm of the ratio of the power spectrum of Y computed considering the possible contribution of X and the power spectrum of Y computed by its own, in both cases evaluated at ω. And it is interpreted as the portion of the power spectrum associated with the residuals that do not take into account the presence of Y [13]. The Geweke spectral measure does not have its equivalent formulation in the information-theoretic framework. As shown in [13], the lack of a temporal separation between the past and the future of the involved processes is what allows a spectral formulation of a parametric criteria. Since in the non-parametric criteria, a way to avoid the temporal separation has not been found up to now, its spectral formulation is not available.

Other examples of causal measures developed in the spectral domain are the Partial Directed Coherence (PDC) [5] and the Direct Transfer Function (DTF) [28]. Both were initially developed under the assumption of identity matrix as covariance matrix of the innovation process and then generalized in [55] where they are named the information PDC (iPDC) and the information DTF (iDTF). Both are defined as a coherence measure between two processes thus they have an interpretation in term of mutual information rate. Moreover, both are measures to test for Granger causality, but only in the case of DTF, a direct connection between the bivariate Geweke spectral measure and the bivariate iDTF exists. iPDC assumes an autoregressive model for the process while iDTF starts with the moving average representation of the autoregressive model.

Especially important in the neuroscience application of these causality measures, is their multivariate extension [39]. In the case of the bivariate iPDC and iDTF, they are straightforwardly extended to the multivariate case [56]. Also, the Geweke measure in time domain has a direct extension of its bivariate formulation done by conditioning on the processes that are not included in the pair [7]. Less immediate is the extension of the spectral representation, for a detailed explanation see [20].

2.3.3 Nonparametric implementation

Regarding the non-parametric approach. Given a criterion of causality, its definition of causal interaction is formulated in terms of identity between probability distributions. Afterwards, a metric is adopted in the information-theoretic framework in order to test whether the identity holds [60, 54].

In the case of the Granger criterion, a widespread measure of its non-parametric implementation is the *transfer entropy*. It is based on a comparison among probability distributions that refer to the hypothesis of independence between the candidate effect and the past of the candidate cause (2.1) [49, 3, 2]. The transfer entropy from the process Y to X is defined as

$$\begin{aligned} T_{Y\to X} &= \sum_{X_{t+1},X^t,Y^t} p(X_{t+1},X^t,Y^t)\log\frac{p(X_{t+1}|X^t,Y^t)}{p(X_{t+1}|X^t)} \\ &= H(X_{t+1}|X^t) - H(X_{t+1}|X^t,Y^t) \\ &= I(X_{t+1};Y^t|X^t) \end{aligned} \tag{2.6}$$

where $H(\cdot)$ indicates the entropy and $I(\cdot)$ the mutual information. In particular, the transfer entropy computes the KL-divergence between the probability distributions $p(X_{t+1},X^t,Y^t)$ and $p(X_{t+1}|X^t)$. By definition, the KL-divergence is non-negative and zero only when the two distributions are equal thus also (2.6) is zero if (2.1) holds. Moreover, the fact that KL-divergence does not consider any specific moments of a given order, is particularly relevant in detecting non-linear interactions. Beyond transfer entropy, other non-parametric measures based on different metrics, have been proposed [4], e.g. the measure based on the Fisher information.

2.4 Machine Learning

In this section we recall some basic concepts of machine learning. In particular, we focus on the problems of *regression* and *classification* since they will be frequently mentioned in the following chapters. Moreover, a short paragraph is dedicated to an evaluation method and to a performance quantification technique. We are referring to the cross-validation technique and the Receiver Operating Characteristic curve (ROC curve) and related Area Under the Curve score (AUC score).

2.4.1 Regression and classification

Both regression and classification are two problems of the so-called *learning theory*. In learning theory, the usual problem is to find a function that well predicts the output variable given the values taken by the input variable. There are three main entities on which we can base our solution: a dataset $\mathcal{D}$, a function class $\mathcal{F}$ and a loss function ℓ. Consider two random variables $\mathbf{x}$ and $\mathbf{y}$, the former is the input variable and it takes values from $\mathcal{X}$, the latter is the output variable and it takes values from $\mathcal{Y}$. Define the dataset $\mathcal{D}$ as

$$\mathcal{D} = \{(x_1, y_1), \dots, (x_n, y_n)\} \sim P^n(\mathbf{x}, \mathbf{y}), \quad x_i \in \mathcal{X}, y_i \in \mathcal{Y} \tag{2.7}$$

the function class $\mathcal{F}$ as a set a functions $f : \mathcal{X} \to \mathcal{Y}$ and the loss function $\ell : \mathcal{Y} \to \mathcal{Y}$ which penalizes the error between the actual output y and its predicted value $f(x)$. The standard approach for solving a learning theory problem consists in finding f in order to minimize ℓ given $\mathcal{D}$. The difference between a regression and a classification problem is in $\mathcal{Y}$. The output y_i is a continuous variable in the case of a regression problem while y_i is a categorical variable in a classification problem. Commonly, to improve the solution, f is not directly applied on the input x_i but a further function g is defined. This function is meant to map the input in a space that facilitates the learning and it is usually known as *basis function* in the case of a regression problem or we refer to the *feature space* as the space in which the input is mapped by g regarding a classification problem. About the notation, we call the set $\mathcal{D}$ data or measured data, and each pair (x_i, y_i) trial or sample. When the problem is a regression one, the input x_i is the regressor variable while the output y_i is the dependent variable. In the case of a classification problem, y_i is the label of the trial x_i.

2.4.2 Evaluation methods

Cross-validation is an approach for having an estimate of the accuracy of a predictive model and gives an idea on how the model will generalize. Beyond an estimate of the model accuracy itself, cross-validation is commonly used as a model selection technique. Cross-validation splits the dataset $\mathcal{D}$ in two disjoint subsets: the training set $\mathcal{D}_{train}$ and the validation or testing set $\mathcal{D}_{test}$. Then, the model is trained using the training data $\mathcal{D}_{train}$ and its accuracy evaluated in the unseen validation data $\mathcal{D}_{test}$. There exist extensions of cross-validation that are meant to give a more robust estimate of the model accuracy. For example, the so-called k-fold cross-validation in which $\mathcal{D}$ is split into k disjoint subsets and cross-validation is repeated k times. At each repetition, the evaluation is done on a different subset and the training on the remaining $k-1$ subsets. Another variant is leave-p-out cross-validation. It uses p samples that are randomly selected from $\mathcal{D}$, as validation set, and the remaining samples as training set. The final accuracy of the model is computed by combining the accuracies of each split.

The Receiver Operating Characteristic (ROC) curve is a graphical plot that shows the performance of a binary classifier under different threshold settings. The curve is done by plotting the true positive rate against the false positive rate. Since many classifiers have a continuous output, in order to obtain the predicted label a further step of discretization should be applied. And this is commonly done by setting a threshold on the real output. Choosing a proper threshold may itself be problematic thus the ROC shows how the classifier performs given various thresholds. The ROC space is a square of side 1, from 0 to 1 on both axises. The best possible classifier would have a point in the upper left corner, i.e. coordinate (0,1). While a completely random guess classifier would plot points along the diagonal from the origin to the upper right corner. This diagonal divides the ROC space into two parts: the upper half in which accuracy is greater that chance level and the lower part in which accuracy is less than chance level. The global performance can be quantified by computing the Area Under the Curve (AUC). Being the ROC space a unit square, in the case of perfect classification the AUC is 1 while it is 0.5 for a random guess classifier.

2.5 Causality in Neuroscience

The application context of this thesis is *neuroscience*. Here we introduce a specific area of neuroscience that goes under the name of *brain connectivity* and in which time series causality is of interest.

2.5.1 Brain connectivity

Brain connectivity aims to investigate the pattern of interactions between distinct units within the brain [27]. The concept of brain units is strongly related to the level of the adopted scale. Thus, brain connectivity can be studied from the microscopic level of single synaptic connections to the macroscopic level of brain regions. Moreover, depending on the type of interactions of interest, brain connectivity is divided into *structural*, *functional* and *effective* connectivity. In the first case, the connectivity patterns are referred to the anatomical links i.e. the neural pathways, in the second case to the statistical dependencies between brain activities in different units and in the last one to the causal interactions between them [44]. In particular, effective connectivity provides information about the direct influence that one or more units exert over another and aims to establish causal interactions among them [17].

2.5.2 Brain signals

Electrophysiological signals are among the most suitable ones for studying effective connectivity. First, because they directly measure neuronal activity, even though at an aggregated level. Second, because their temporal resolution is compatible with the processing time at the neuronal level, that is in the order of milliseconds [48]. These data can be measured with invasive or non-invasive methods. Invasive methods allow a high quality and spatially precise acquisition by implanting electrodes on the brain. On the side of the non-invasive techniques, magneto- and electro-encephalography (M/EEG) are widely used because they directly measure neural activity with a high sampling frequency and, by means of source reconstruction techniques, they provide increased signal-to-noise ratio and spatial resolution [11].

Another well established technique to study brain activity is the functional magnetic resonance imaging (fMRI). We do not enter in the details of the generation of the signal, we only mention that the effect captured by the machine is the variation of the concentration of deoxyhemoglobin within tissues. In particular, the origin of this variation is that an increase of the neural activity implies an increase in the local blood flow. And more importantly, deoxyhemoglobin is paramagnetic thus it can be detected by the scanner [21]. Differently from the previous measuring techniques, fMRI is an indirect measure of the brain activity since the physical phenomenon acquired by the machine is a consequence of the quantity of interest. Typically, the relationship between the measured (BOLD) signal and the underlying brain activity is modeled by a linear time invariant (LTI) system in which the BOLD signal is the result of the actual brain activity with the so-

called hemodynamic response function (HRF). This filtering effect of HRF on the actual brain activity is a key point when BOLD signal is used to study causality. Indeed, how HRF affects the inference of time lag-based methods, e.g. Granger-based methods, is an open problem. There are studies that state that the BOLD signal is not compatible with the assumptions of precedence and predictability that are at the root of Granger causality [15, 47], while others prove the robustness of Granger causality to variations of HRF and identify other factors, e.g. SNR and time resolution, as potential issues in causal inference [51].

2.5.3 Causal connectivity

The interest in studying causal interactions from neuroimaging data is not only limited to effective connectivity but it has a more general scope. The original definition of effective connectivity provided in [17], refers to the directed influences that neuronal populations in one brain area exert on those in another one. Thus an estimator of effective connectivity should consider the physiological structure and dynamics of the system [18]. This constraint is particularly demanding since it means modeling the underlying physical processes. To overcome this issue, a relaxed version of effective connectivity was introduced in [10] under the name of *causal connectivity*. Causal connectivity refers to a causality measure that infers the causality structure without requiring it to be representative of the underlying neuronal network.

The term *causality analysis* is commonly used when studying the direct interactions among brain signals. As highlighted in [14], a causality analysis may have different meanings. Its purpose could be to infer the existence of a direct causal connection, thus the estimate of the so-called causal structure or (binary) causal graph [16]. A different goal is to study the mechanism underlying a causal connection. This means focusing on how a causal connection is physiologically implemented. And a third question concerns the quantification of the interaction, thus it requires both an appropriate modelling of the dynamics and a clear understanding of what the causal effect coming from the causal connection, actually means [46].

Chapter 3

Problem Statement

In this chapter, we present the three problems addressed by our work. From a general point of view, all the problems are positioned in the area of causal inference among time series, i.e. time series causality. More precisely, the first problem regards the multivariate implementation of the Granger criterion in the standard unsupervised approach. Then in the second problem, we consider a different strategy for tackling the problem of time series causality and we propose an approach that lies in the area of machine learning. We will refer to it as the supervised parametric approach. The last problem concerns the parametric implementation of the proposed supervised approach. In particular, our interest is on the effect of using a neuro-physiological model instead of the standard MAR model when describing the stochastic process of data generation.

3.1 Bayesian model identification with structured prior

In the literature on causal inference, and specifically on the family of the parametric methods implementing the Granger criterion of causality, a common step is the identification of the MAR model. In practice, this means to estimate the coefficients of the model as well as the residual covariance matrix. There are two main approaches to estimating the MAR model: by solving the Yule-Walker equations, or by applying a linear regression estimator. The former approach has been largely applied, especially in the past, and it requires the solution of a system of linear equations. Solving the Yule-Walker equation leads to a stationary model that can be solved iteratively [40]. On the other hand, in [30] it has been shown that the latter approach should be preferred. The application of a linear regression method allows more flexibility in the inference process since it can better deal with the finite number of samples, the potential high dimensionality of the problem

and the risk of over-fitting.

In order to state the problem of MAR identification in terms of linear regression, we rewrite Equation 2.3 as

$$\mathbf{y}_t = \sum_{i=1}^{p} \mathbf{A}_i^{\mathrm{T}} \mathbf{y}_{t-i} + \mathbf{e}_t \,, \tag{3.1}$$

where $\mathbf{y}_t$ denotes a $d_y \times 1$ vector, representing the state of d_y time series measured at time t, and p is the order of model. Moreover, $\mathbf{e}_t \sim \mathcal{N}(\mathbf{0}, \mathrm{diag}(\sigma_1^2, ..., \sigma_{d_y}^2))$ is the so-called innovation process, with temporally independent increments for which each time instant is a realization from a d_y-dimensional Gaussian distribution with zero mean and diagonal covariance matrix. The $\mathbf{A}_i \in \mathbb{R}^{d_y \times d_y}$ with $i = 1, 2, \ldots, p$ are the coefficient matrices that model the influence of the signal values at time $t - i$ on the current signal values at time t. $\mathbf{A}_i$ matrices are derived from the *causal configuration matrix* $\mathbf{A}$, that represents what a time series causality method should infer given $\mathbf{y}$.

The so-called standard form of the model can be easily derived by constructing the $(d_y p) \times 1$ vector $\mathbf{x}_t = [\mathbf{y}_{t-1}^{\mathrm{T}}, \mathbf{y}_{t-2}^{\mathrm{T}} \ldots \mathbf{y}_{t-p}^{\mathrm{T}}]^{\mathrm{T}}$. $\mathbf{x}_t$ contains the past dynamics of each time series needed to compute the current amplitude $\mathbf{y}_t$. All the $\mathbf{A}_i$ coefficient matrices of each time lag are vertically stacked in a unique $(d_y p) \times d_y$ matrix $\mathbf{W} = [\mathbf{A}_1; \ldots; \mathbf{A}_p]$. Thus

$$\mathbf{y}_t = \mathbf{W}^{\mathrm{T}} \mathbf{x}_t + \mathbf{e}_t \,, \tag{3.2}$$

which shows that the model can be identified by solving a multivariate linear regression problem.

The problem of fitting a certain type of curve to a set of measurements has been largely studied in science since centuries [35]. And as a consequence, the related literature is very large. Here, we focus on two relevant aspects for our domain of application: the high dimensionality of the problem and the availability of prior knowledge that may improve the inference. The first aspect refers to the nature itself of neuroscientific datasets since generally a large amount of data are collected from multiple brain regions. Thus, there is the need of regularizing the inference to overcome issues like the over-fitting and the non-uniqueness of the solutions. The second point is related to the possibility of including prior knowledge on the causal structure. In order to address these two elements, a Bayesian-based method appears to be a meaningful approach. Indeed, in the Bayesian setting, regularization can be interpreted as imposing a particular prior on the model coefficients. Moreover, the Bayesian inferential process allows the inclusion of prior knowledge in the model definition, the use of the model evidence as a measure to compare hypotheses and

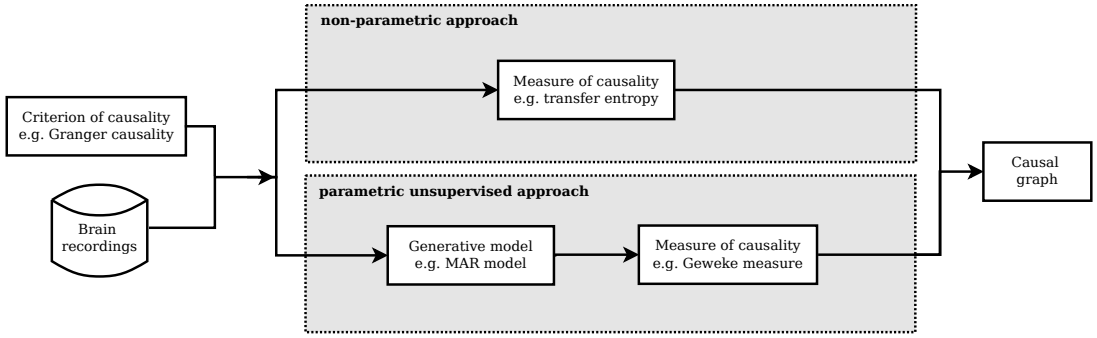


Figure 3.1: Given a criterion of causality, the estimation of causality structure can be mainly implemented in two different ways: the non-parametric approach (top) and the parametric one (bottom).

also a quantification of the residual uncertainty as captured by the posterior distribution.

In particular, our interest is in the so-called structured prior. This type of prior follows the idea of grouped variables in the sparse regularization methods. Thus, it allows the definition of different groups of coefficients that are separately regularized. Prior knowledge on the causal structure is included by grouping coefficients that are supposed to be drawn from the same prior distribution.

Summarizing, the purpose of this first problem of interest for this dissertation is to define and study a Bayesian approach with a structured prior to solve a multivariate linear regression problem in the context of Granger causality for time series.

3.2 From a criterion of causality to a causal graph

According to the literature that was revised in Section 2.3, there are two main ways to obtain an estimate of the causal graph associated to a set of time series, see Figure 3.1. The so-called non-parametric approach, which is based on defining a measure in the framework of the information theory, and the parametric approach, in which the measure is defined on a specific generative model. Here, we follow the parametric approach, and the problem of which we aim to provide a solution is the inference of the causal graph given both a criterion of causality and a stochastic generative model. As causal graph, or causal configuration matrix, we refer to a $n \times n$ binary matrix that indicates the causal interactions among a set of n time series.

By definition, in any parametric approach a generative model is assumed as representative of the stochastic process. In the domain of causal inference, the key element that

the model needs to implement is the concept of causal link between time series. The definition of causal link derives from the chosen criterion of causality and this latter is not necessarily application dependent. An example is the Granger criterion and the related MAR implementation. Indeed, they are not constrained to a specific type of application, as suggested by their wide application in many fields of science. And this is due to the capability of the MAR model to suit for a large variety of time series, as well as for electrophysiological signals [24]. However, the MAR model does not consider the physiological mechanism of data generation. This on one hand, reduces the complexity of the model and makes it feasible to invert but, on the other hand, it doest not allow an interpretation of the inference in terms of effective connectivity.

In the area of computational neuroscience, a large body of literature has been developed on realistic model of neural processes. Despite the level of realism of the models, their inversion and inference remain an open problem. As an example, we mention the Dynamical Causal Modeling (DCM) [34]. DCM overcomes the physiological plausibility issue by adopting a more realistic generative model. It considers the dynamics exhibited among neural populations by the definition of a forward model in which hidden states are mapped to observed quantities. Due to its complexity, i.e. a large number of free parameters, DCM is not an exploratory technique. Meaning that it does not infer the causal graph of a given set of signals. DCM estimates the posterior probability of a specific hypothesis given the data, thus it is used as an approach to test predetermined hypotheses and to discover the most likely one.

Furthermore, a crucial aspect of any Granger-based method concerns the bivariate comparison on which the evaluation is based. More precisely, according to the definition of Granger causality, each pair of observed variables is separately evaluated in order to determine the presence of a causal link. Only afterwards, the conditional version has been introduced, in order to consider the other variables under analysis but without changing the bivariate nature of the evaluation. This type of approach does not address the multivariate aspect of the data.

As an attempt to overcome these issues, we propose an alternative approach to the traditional unsupervised one. Our proposal is based on the concept of learning the causal structure from examples. In other words, the proposed approach lies in the area of supervised learning methods. We will refer to it as the parametric supervised approach. A detailed description of the proposed approach is given in Section 4.2.

The basic idea of the supervised approach is to unveil causal connections through a classification schema. The main components are i) the training dataset, also called

representative dataset, generated by the adopted generative model, ii) the feature space, that is defined according to a specific criterion of causality and iii) the classification schema that allows the prediction of the causal graph of the brain recordings. In order to make the supervised approach comparable with the standard parametric unsupervised one, we firstly evaluate it in the case of the MAR implementation of the Granger causality. Thus, referring to the notation introduced in Section 2.4, the problem is decomposed into subproblems that need to be separately faced in order to define the supervised approach.

- $\mathcal{D} = \{(x_i, y_i)\}$ where x_i is a realization of a M dimensional MAR process, see Subsection 4.2.3;
- $y_i \in \mathcal{Y}$ where $\mathcal{Y}$ can be $[0, 1]$ or $[0, 1, \ldots, 2^{M(M-1)} - 1]$, it depends on the classification schema, both will be considered see Subsection 4.2.2;
- $g(\cdot)$ regards the definition of the feature space that will be defined according to the MAR implementation of the Granger criterion by considering the multivariate nature of the input signals, see Subsection 4.2.1;
- a k-fold cross-validation framework is adopted for the evaluation part.

Each of the listed points will be instantiated in the next chapter according to the scope of each experiment.

3.3 The role of the generative model in the parametric supervised approach

As we mentioned in the previous section, the generative model beyond to implement the criterion of causality, it should also consider the specific scenario of application. Since this work is meant to be applied in the context of neuroscience, we investigated under specific assumptions, whether and how considering the domain of application improves the causal inference.

More in details, our question is whether the supervised approach can be improved when a more plausible model of the specific context of application is available. Considering our field of application, this investigation goes in the direction of making the inference more interpretable and so moving from a causal connectivity analysis to an effective connectivity analysis.

Recent progresses have been done in neural network modeling and they make possible to generate models with biophysical and anatomical properties very similar to those of

real cortical circuits. Moreover, if these models are simulated as dynamical system, the generated activity has been shown to have statistics very close to that of recorded cortical activity. We aim to take advantage of the availability of these techniques by adopting this type of generative model that we call neural network (**NN**) model, and generate a **NN** dataset to study how the parametric causal inference performs on it.

Firstly, we will analyze the problem of directly applying the supervised approach on the **NN** dataset:

- $\mathcal{D}_{train} = \{(x_i, y_i)\}$ where x_i is a realization of a M dimensional MAR process;
- $\mathcal{D}_{test} = \{(z_i, y_i)\}$ where z_i is a realization of a M dimensional **NN** model;
- $\mathcal{Y}$ is kept the same across datasets;
- $g(\cdot)$ regards the definition of the feature space that will be defined according to the MAR implementation of the Granger criterion by considering the multivariate nature of the input signals, see Subsection 4.2.1;
- for the evaluation a similar k-fold cross-validation approach is used in order to make results comparable although there are two different datasets for training and evaluating.

Then, the **NN** model will be used also as generative model of the training dataset thus as model assumed by the supervised approach as representative of the process of data generation. The problem is structured as follows:

- $\mathcal{D} = \{(x_i, y_i)\}$ where x_i is a realization of a M dimensional **NN** model;
- $\mathcal{Y}$ is determined by the classification schema;
- $g(\cdot)$ as before, so based on the MAR implementation of the Granger criterion;
- a k-fold cross-validation framework is adopted for the evaluation part.

Chapter 4

Solutions

In this chapter, we will recall the problems presented before and a solution is proposed to each of them. More in details, in Section 4.1 a solution to the identification of a linear model is proposed in the context of the Bayesian inference with the possibility of defining constraints on the sparsity structure of the independent variables. Section 4.2 is dedicated to the problem of inferring the causal graph from a given set of time series. And a supervised parametric approach will be presented and customized in the context of the Granger criterion of causality. Finally, Section 4.3 focuses on a specific aspect of the supervised approach and its purpose is to study whether a more detailed generative model with respect to the context of application, may improve the inference.

4.1 Bayesian approach for linear model identification with structured prior

In Section 3.1 we saw that the problem of identifying a MAR model can be formulated in terms of a linear regression problem. Many solutions have been proposed starting from the simple minimization of the root mean square error to more sophisticated penalized regression model. In particular, a large body of literature concerns the so-called *group sparse regularization* methods. This represents a sort of extension and generalization of the concept of regularization. Regularization was introduced to overcome limitations of the data fitting methods [45] such as overfitting, non-unique solution, high correlation between signals etc [25, 57]. It consists in including in the cost function the so-called penalty term that controls the overall amplitude of the estimates. Regarding the group sparse regularization methods, they are based on the idea of *grouped variables*. By grouped variables, it is meant that the independent variables are clustered to allow a selection of

explanatory variables that better explain the output. In other words, instead of regularizing the whole coefficient vector at once, each cluster is separately regularized. In the Bayesian setting, regularization can be interpreted as imposing a particular prior called *structured prior* on the model coefficients. The concept of structured (or group sparsity or sparsity-enforcing) priors conveys the same idea of grouped variables. Thus, a structured prior refers to a clustering of the coefficients in which elements in the same group are drawn from the same prior distribution.

4.1.1 Gaussian scale Mixture Expectation Propagation (GMEP) method

We propose a novel approach for Bayesian group sparse modeling. We will refer to the model as GMEP. The name refers to the Gaussian scale Mixture distribution that is used to form a general class of group sparsity priors, and to the Expectation Propagation framework that is used for approximating the inference. For a more extensive treatment of the method, we refer to Chapter 7. Here, a general description is given by Figure 4.1 which shows the graphical representation of GMEP. Circles represent random variables, while rectangles denote known variables. The fixed hyperparameters $\boldsymbol{\mu}_{\theta,0}$ and $\boldsymbol{\Sigma}_{\theta,0}$ are denoted with dots. Considering the notation used in Section 3.1, $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_n]^{\mathrm{T}}$ is the $n \times d_y$ output variable matrix, $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_n]^{\mathrm{T}}$ is the $n \times d_x$ input variable (or design) matrix, and $\mathbf{W} = [\mathbf{w}_1, ..., \mathbf{w}_{d_y}]$ is the $d_x \times d_y$ coefficient matrix. $\boldsymbol{\theta}$ is the hyperparametr vector and it contains both the hyperparameters of the likelihood terms and the ones of prior terms. $\boldsymbol{\theta}$ is modeled by a fixed multivariate Gaussian prior density with hyperprior mean vector $\boldsymbol{\mu}_{\theta,0}$ and hyperprior covariance matrix $\boldsymbol{\Sigma}_{\theta,0}$. $\mathbf{W}$ is modeled by a structured Gaussian scale-mixture prior and a Gaussian observation model is used for each output variable. $\mathbf{V}$ and $\mathbf{U}$ are known transformation matrices that select the desired model parameters according to a specific time point and output variable. In particular, $\mathbf{V}$ is used for selecting the variance hyperparameter of both the likelihood and the prior distributions, while $\mathbf{U}$ is used for defining the structured priors.

4.1.2 Employed structured coefficient priors

Since a large part of this activity is concerned with the characterization of the model under different structured priors and datasets, we give here an intuitive description of the three structured priors adopted in the experiment part. Firstly, a *uniform Gaussian prior* was defined for each output. Such configuration is strictly related to ridge regression because the coefficients associated with each output are supposed to belong to the same group. That means they are modelled as drawn from the same distribution. In other words,

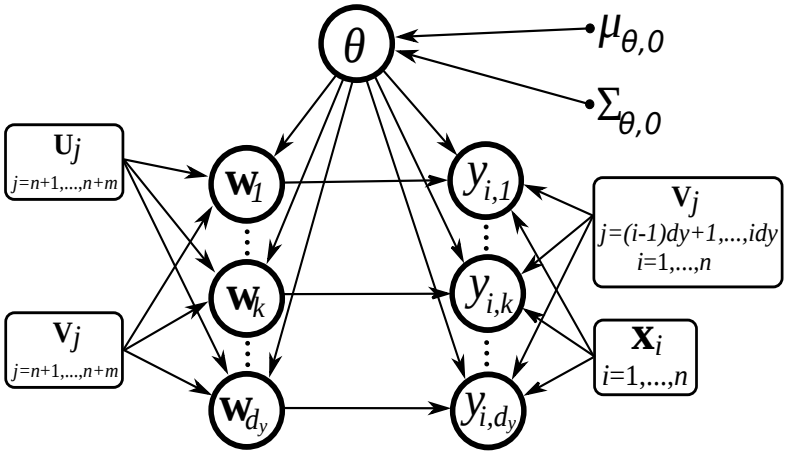


Figure 4.1: Graphical model of GMEP in which dependences between variables are shown by using circles for random variables, rectangles for known variables and dots for fixed hyperparameters.

we can see this as the GMEP implementation of ridge. The second trivial configuration that was taken into account, considers one group for each coefficient. This represents the opposite situation with respect to the previous prior, thus now each coefficient has its own distribution to which it belongs to. This approach is known as *automatic relevance determination* (ARD) because the hyperparameters of each distribution determine the sparsity i.e. the relevance, of the related coefficient. The third case in our comparison has a definition of groups that reproduces the true sparsity structure of the coefficients in $\mathbf{W}$. Considering the null hypothesis of zero causality, i.e. $a_{x,y,i} = 0, \forall i$, we see that the same sparsity structure is shared across time lags, i.e. the amount and position of the zero connections are the same across $\mathbf{A_i}$. This can be rephrased as: the causal configuration is time independent, i.e. there is no dynamic in the causal interactions Therefore, we call that prior group the *lag-independent prior*.

4.2 Supervised causal inference

With regard to the problem of time series causality, we formulate it in order to be solved in the framework of the supervised machine learning.

The idea of posing causal inference as a learning theory problem is not new. An example is [29] where the authors adopted a supervised approach for bivariate causal inference with the use of kernel mean embeddings for feature mapping.

Here, the same idea of a supervised detection of causal interactions is used but with a

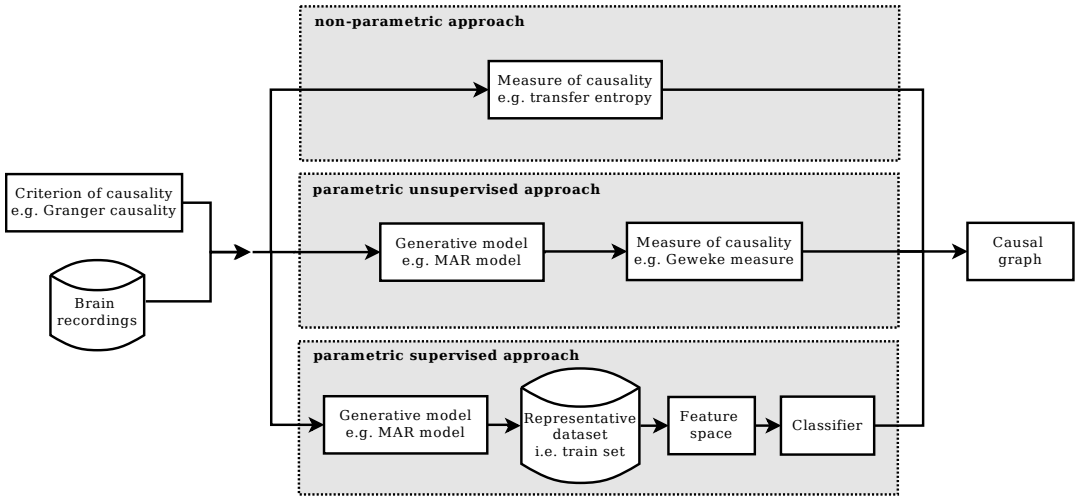


Figure 4.2: Given a criterion of causality, the estimation of causality structure can be implemented in three different ways: the standard non-parametric approach (top), the parametric one (mid) and the proposed parametric supervised one (bottom).

different implementation and specifically contextualized for time series causality. In our version, the model is not used to derive a measure but to generate a dataset that is meant to represent the population of causal graphs of interest. The purpose of this dataset is to be used as train set of a standard classifier, aimed to predict the causal graph of future multivariate time series. A consequence of the proposed approach is that we need to build a feature space in which to represent the dataset. And the definition of the feature space is directly connected with the chosen causality criterion. Indeed, the role of the feature space is to implement the criterion of causality in order to encode the causal structure exhibited by the trial. Moreover, it is interesting to notice that model and feature space do not need to derive from the same causality criterion. This means that the proposed approach allows to disentangle the mechanism of data generation from the criterion used to describe the causal structure. Figure 4.2 shows the parametric supervised approach compared with the other two that we discussed in Section 3.2.

The analysis that we conducted on the supervised parametric approach is based on the Granger criterion of causality. Similarly to what was done in the activity with GMEP, the Granger criterion was implemented by the MAR model. And also the feature space was defined according to the same implementation.

4.2.1 Definition of the feature space

The feature space directly considers the MAR implementation of the Granger criterion. As we have seen in Section 2.3.2 the bivariate conditional (or simply the bivariate) Geweke measure in time is based on a comparison between two residual variances. Relying on the same idea, a given trial of M time series is mapped in a vector of measures that quantifies the ability to predict a certain time series (effect) from the past values of each possible combination of causes. All the possible combinations of causes are the subsets that can be defined from M time series thus $\sum_{i=1}^{M} \binom{M}{i}$. By considering that at each combination of causes, one of the M effects can be assigned, the total number of pairs causes/effect is $\sum_{i=1}^{M} \binom{M}{i} M = (2^M - 1)M$ (by using the binomial theorem). We refer to each of these pairs as a *causality scenario*. Table 4.1 shows the causality scenarios when $M = 3$. For each

	Causes	Effect
1	x_0	x_i
2	x_1	x_i
3	x_2	x_i
4	x_0, x_1	x_i
5	x_0, x_2	x_i
6	x_1, x_2	x_i
7	x_0, x_1, x_2	x_i

Table 4.1: For each effect $x_i, i = 0, 1, 2$ and $M = 3$, we report the 7 possible causality scenarios.

causality scenario, a plain linear regression problem is built by selecting, as dependent variable, the time points from the signal in the *effect* column. Each of these dependent variables has a regressor vector composed of the p previous time points selected from the signals in the *causes* column, where p is the order of the MAR model. Finally, the regression problem of each causality scenario is scored, by common metrics like the means squared error. Such scores are used as features in the feature space representation.

As summarized in Table 4.1, the feature space is defined by exploiting all the possible causality scenarios among a set of M time series. We call it the *complete* feature space. But it is also worthwhile to look at the bivariate Geweke measure in time in terms of causality scenarios. This allows the definition of a feature space that is based on the same information that would be used in the standard unsupervised parametric approach. In the bivariate case, given M time series and selected one as effect, its possible causes (that are $M - 1$) define $M - 1$ causality scenarios plus the causality scenario of the reduced univariate representation. Thus, in total for each effect, we evaluate M causality scenarios.

This means that a feature space based on the pairwise Geweke measure has M^2 features and we will refer to it as the *pairwise* (pw) feature space. Similarly, by repeating the same reasoning with the conditional pairwise measure, the only difference is in the subset of causes since now it contains also the $M-2$ time series that are not in the pair under analysis, i.e. the variable Z in Equations 2.3 and 2.4. We will refer to this feature space as the *conditional pairwise* (c-pw) feature space.

4.2.2 Classifications schema

Here, we describe two versions of the parametric supervised method. In the first, the entire causal configuration matrix $\mathbf{A}$ is considered the class label of the trial. This choice implies that one classifier has to be trained to discriminate among $2^{M(M-1)}$ classes. We will refer at this solution as the *matrix-based classification* (MBC). In the second version of the parametric supervised method, each cell of the configuration matrix is analyzed independently from the others. Since each cell can be only 0 or 1, then the whole problem of predicting the causal configuration is transformed into $M(M-1)$ binary problems, one for each cell. We call this approach the *cell-based classification* (CBC).

4.2.3 Representative dataset and analysis of the method

Regarding the generative model, the MAR model was used to generate a class-labeled dataset of which we refer to as $\mathbf{L}$. More precisely, the generative model was not as described in the Equation 3.1 but a variation of it that includes an additive noisy component. $\mathbf{L}$ is generated considering the total number of causal graphs that can be produced by M time series and it will be the *representative dataset* used as train set. In the causal configuration matrix $\mathbf{A}$ there are $M(M-1)$ free binary parameters and so $2^{M(M-1)}$ possible causal configuration matrices. Considering that $\mathbf{L}$ must be representative of the entire population of configurations, it is generated so that multiple trials are included for each possible causal graph.

Concerning the analysis part, we performed experiments on both synthetic and real data. The purpose of the experiments with synthetic data, i.e. $\mathbf{L}$ dataset, is to compare the proposed supervised methods against the standard conditional pairwise Geweke measure in time. We used the method proposed in [6] as (unsupervised) implementation of the Geweke measure and we refer to it as the GCA method. Moreover, the supervised approach was evaluated across different feature spaces, i.e. the complete version, pw and c-pw. Additionally, on the real data, we investigated the behavior of the supervised approach when the underlying exact generative model is not known in advance.

4.3 Neurophisiological modelling of brain signal for supervised causal inference

Referring to Figure 4.2 and as we mentioned before, the generative model in the supervised approach may be not related to the adopted criterion of causality. In other words, the representative dataset in which the classifier is trained, may not be consistent with the criterion of causality used for the definition of the feature space. Since the purpose of this third activity is to investigate how a more plausible generative model with respect to the context of application, affects the inference, we used a representative dataset that is as similar as possible to the physiological recordings.

We decide to use a model based on [31] in which a cortical network model composed of leaky integrate and fire neurones, allows a behavior that strongly resembles the primary visual cortex. By connection several of these cortical network models, we can generate a simulated information flow among neural circuits that has realistic statistical properties and for which we know the ground truth of causal configuration. Each simulated network is composed of 5000 neurons of which 80% are taken to be excitatory and the remaining 20% are inhibitory. An inter-network directed connection is established by linking 20% of the pairs composed of any cell from the receiver network and an excitatory cell from the sender network. A detailed description of the model is given in Chapter 9 Subsection II-B. We refer to the dataset generated by this model as the **NN** dataset.

We evaluate the capability of inferring the causal graph in **NN** both when the supervised method (CBC) assumes the MAR model as generative one, and also when the physiological model is chosen for the generative process. While regarding the feature space, we always use the complete one as it is described in the Section 4.2.1. We compare these two scenarios with the unsupervised inference done by GCA.

Chapter 5

Results

The results collected from the analysis of the proposed solutions, are presented in this chapter.

5.1 Evaluation of GMEP based on the structured prior

Starting from the activity related to the GMEP method, we firstly recap its main purpose: investigate how the structured prior affects the final inference in a simulation framework. Specifically, we aim to show that a more detailed modeling of the group sparsity prior, through the inclusion of information related to the structure of the data, improves the inference of GMEP. A comparison across three different priors is conducted, and we refer to the structured priors in Section 4.1.2. Results are reported by using the uniform Gaussian prior as baseline with which the other priors are compared. The predictive performances are evaluated by computing the mean log predictive density (MLPD). For a detailed explanation of the measure, refer to Chapter 7 Subsection IV-A. Higher MLPD values corresponds to higher approximate predictive density values for test data points on average indicating better predictive performance. Moreover from the EP iterations, an approximation of the mean leave-one-out predictive density can be derived. We refer to it as MLPD_{EP} and we use it as an estimate of the future predictive performance of the model. A third variation of MLPD is computed if the ground truth of the coefficient vector is available. We call it $\text{MLPD}_{\mathbf{w}}$ and it measures how well the posterior approximation matches with the true coefficients. These three measures are respectively reported in Figures 5.3, 5.2 and 5.1.

The second experiment related to GMEP is on an empirical fMRI dataset. A detailed description of the dataset is given in Chapter 7 Subsection III-B. The purpose of this

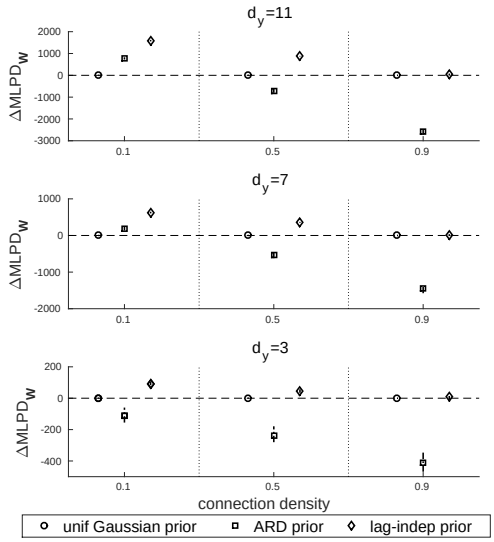


Figure 5.1: ΔMLPD computed with respect to the uniform Gaussian prior and evaluated on the coefficient estimates.

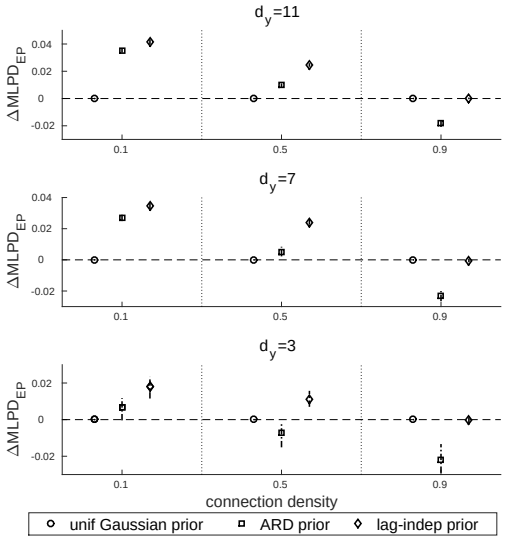


Figure 5.2: ΔMLPD computed with respect to the uniform Gaussian prior and evaluated on the EP iterations.

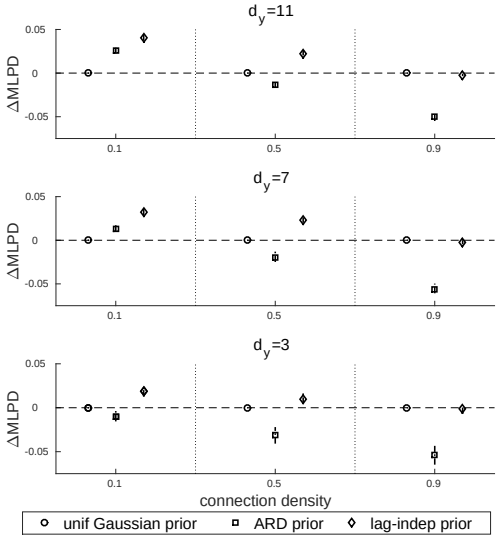


Figure 5.3: ΔMLPD computed with respect to the uniform Gaussian prior and evaluated on the test set.

experiment is to use GMEP to test hypotheses about the sparsity structure of the signals and to exploit the possibility of including prior knowledge of the dataset in the group sparsity prior. In this example, we assume a difference in the magnitude of the coefficients that connect areas in the same hemisphere with respect to the ones that connect areas across hemispheres. We encoded this assumption in a specific group sparsity prior. In details, firstly GMEP was applied using the three structured priors that we adopted also in the simulated dataset. Then, we considered the anatomical position associated with each time series, thus we enriched the three initial priors by adding four new groups in which the coefficients have been clustered according to the hemispheres that they link with. The results of these two scenarios, i.e. the three original structured priors and their extension with anatomical information, are compared with the prior that only models the hemisphere structure. Summarizing, in total there are 7 different structured priors and the comparison is done for different time series lengths, Figure 5.4. This allows an evaluation of the inferences under both different structured priors and number of training time points (the testing set is kept constant).

5.2 Analysis of the parametric supervised approach

Regarding the supervised approach for time series causality, we report here two groups of experiments. In the first group, the generative model used for the representative dataset

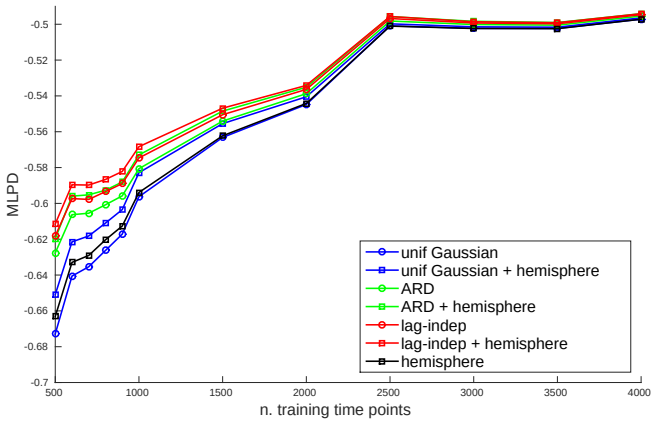


Figure 5.4: MLPD on the test set computed by multiple applications of GMEP under differently structured priors and by varying the number of time points in the training set.

is exactly the same of the dataset to be tested. Both the unsupervised (GCA) and supervised methods are applied in the **L** dataset and the related results are shown in Table 5.1. For the supervised approach after the mapping of the dataset to the feature space, the logistic regression classifier with l_2 regularization was applied in a 5-fold cross-validation framework. The table quantifies the inference performances in terms of ROC AUC. In particular, we notice that the AUC score changes from 0.72 for GCA to 0.90-0.92 for the supervised methods.

The outcomes of a second type of comparison are shown in the same table, we refer to a comparison between different feature spaces for the supervised approach. More in details, the AUC of the complete feature space, i.e. columns CBC and MBC, the pairwise one, i.e. CBC pw, and the conditional pairwise, i.e. CBC c-pw, are reported. Of the same experiment, the related ROC curves are shown in Figure 5.6.

A third comparison that we can make by looking at Table 5.1, is between the approaches of classification, i.e. the cell-based (CBC) and the matrix-based (MBC) approaches. Columns 2 and 5 report the AUC of respectively CBC and MBC together with that of GCA in column 1. The related ROC curves are represented in Figure 5.5. It is worthwhile to notice that the score of GCA when applied in **L** does not allow a false positive rate lower than 0.55.

The second group of experiments aims to investigate the supervised approach when a mismatch is introduced between the generative model of the representative (training) dataset and the actual process of signal generation. This is the common scenario in the practical case because generative models are only approximations of the real physical

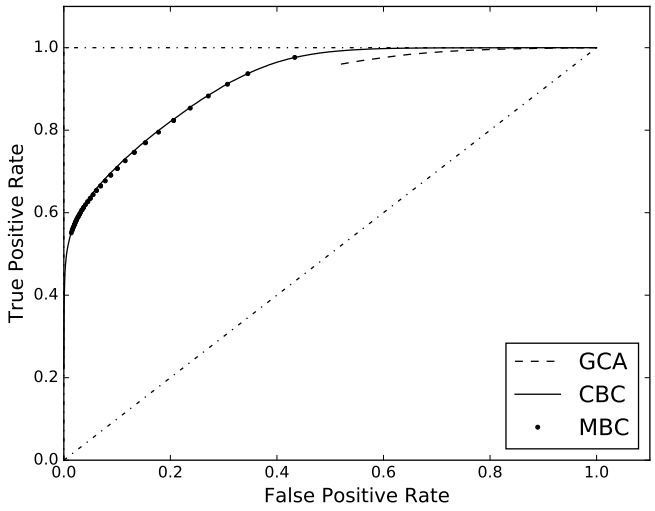


Figure 5.5: ROC curves estimated on the results of the three analysed causal inference methods: Granger Causality Analysis (GCA), Cell-based Classification (CBC) and Matrix-based Classification (MBC).

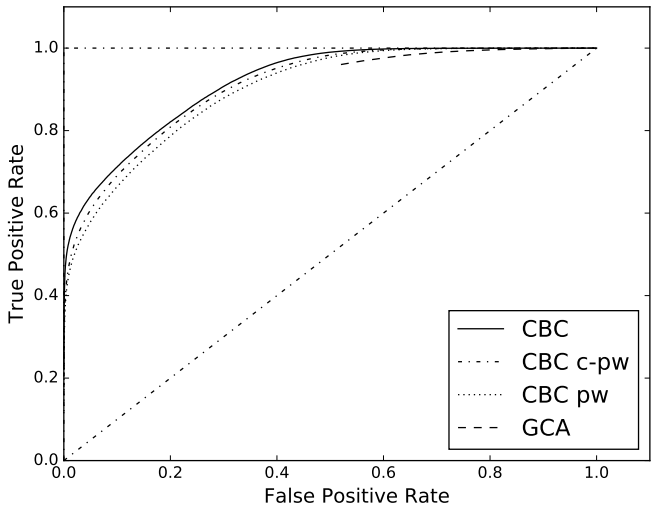


Figure 5.6: ROC curves estimated on the results of CBC when applied on three different feature spaces: the complete one in contrast with the pw and c-pw ones. The ROC curve of GCA is shown as benchmark.

	GCA	CBC	CBC c-pw	CBC pw	MBC
L dataset	0.72	0.92	0.91	0.90	0.91

Table 5.1: AUC values related to the application of GCA, CBC (also with the reduced feature spaces) and MBC on the **L** dataset.

process.

Firstly, before the application on real data, the inconsistency between representative and validation datasets was simulated by generating a new dataset that we name $\mathbf{L}_{\text{MAR}}$. $\mathbf{L}_{\text{MAR}}$ differs from **L** only on the noisy component which is absent. The effect of the additive noise on the inference was evaluated by training CBC in $\mathbf{L}_{\text{MAR}}$ and then applying it in **L**. The resulting AUC is 0.85.

Then the supervised method was applied in a real dataset after being trained on **L**. In particular, CBC was applied on the neural recording dataset of which we expect to infer a specific causal graph given by previous studies [59]. For a detailed description of the dataset, refer to Chapter 8 Subsection II-B. The causal graph was repetitively inferred with different dataset configurations, i.e. the inference was done under different combinations of sampling frequency and time window width. As sampling frequency, we set it to 600, 800 and 1000 Hz and the model order was computed in order to have time windows of 5, 10, 15, 20 and 25 ms. For each pair of sampling frequency and model order, the AUC was computed using as true causal configuration matrix the causal chain EC3→CA1→EC5. Results are shown in Table 5.2.

	5ms	10ms	15ms	20ms	25ms
600Hz	0.80	0.82	0.82	0.83	0.82
800Hz	0.82	0.82	0.82	0.73	0.62
1kHz	0.82	0.82	0.75	0.61	0.64

Table 5.2: AUC computed by applying CBC to the empirical dataset with different sampling frequencies and time window widths.

5.3 Effect of a physiologically plausible generative model

In the previous example, we saw a real application of the supervised approach in which both the generative model and the feature space rely on the MAR implementation of the Granger criterion. This corresponds to assuming that the MAR model is a good approximation of the stochastic process underlying the validation dataset. Here, we used a more realistic model for generating a new dataset, called **NN** dataset. For a detailed

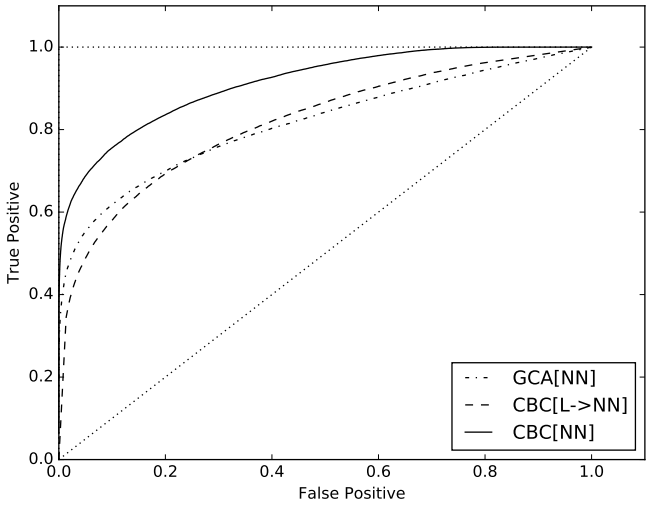


Figure 5.7: ROC curves from the application of GCA and CBC on the **NN** dataset. CBC is applied twice with different training phases. CBC[$\mathbf{L} \rightarrow \mathbf{NN}$] indicates that the method was trained on **L** while for CBC[**NN**] the training was done directly on the **NN** dataset.

description of the dataset generation, refer to Chapter 9 Subsection II-B.

On **NN** we run two experiments. Firstly, causality was estimated using GCA and CBC. The application of CBC was set in order to run under the same condition of GCA. This means that the training was done on the **L** dataset and the feature space was the complete one. Results are shown in Table 5.3 in which the ROC AUCs are reported and in Figure 5.7 that shows the ROC curves.

The second experiment wants to exploit the possibility offered by the supervised approach to disentangle the generative model from the criterion used to decode a causal interaction. To evaluate this scenario, we inferred the causality in the **NN** dataset by training CBC on a feature space whose features are still Granger-based, but it is constructed on a representative dataset generated by the neuro-physiological model. In Table 5.3 and Figure 5.7 this last experiment is indicated as CBC[**NN**].

	GCA[**NN**]	CBC[$\mathbf{L} \rightarrow \mathbf{NN}$]	CBC[**NN**]
AUC	0.82	0.82	0.91

Table 5.3: AUC values related to the application of GCA and CBC on **NN**.

Chapter 6

Discussion and conclusion

This final chapter concludes the first part of the thesis. The main results that have been described in Chapter 6 are now discussed. Moreover, Section 6.2 concludes our work and some possible future developments of these activities are listed in Section 6.3.

6.1 Discussion

GMEP on the simulated dataset Through the analysis presented in Section 5.1 we characterized a novel approach for Bayesian linear modeling with structured prior (GMEP). Our goal is to apply it in the context of MAR identification as initial step of a Granger-based estimate of the causal brain connectivity. One of the main advantages of GMEP is its flexibility in the definition of the structured prior. In order to better understand this property, we designed a simulation study to test how the structured prior affects the prediction under different conditions of dimensionality and connectivity density, i.e. sparsity. We modeled the sparsity by two types of structured priors: the ARD prior and the lag-independent prior. Both are compared with the uniform Gaussian prior.

The lag-independent prior models the actual sparsity structure of the coefficients since it is designed according to the assumptions of the MAR model thus forming an optimal compromise in terms of model complexity. By looking at Figures 5.1 and 5.3, we see that the lag-independent prior always outperforms the other priors or, in the worst case, it is equal to the uniform Gaussian prior.

The uniform Gaussian and the ARD priors can be seen as two extreme cases in terms of model complexity. Regarding the uniform Gaussian prior, the model complexity is very low since all the coefficients that are involved in the modeling of the same time series, are clustered in the same group. Thus, they are supposed to be drawn from the same

distribution, i.e. they are assumed to share the same sparsity level. And by considering our experimental design this assumption is realistic only in the case of very high connection density. Indeed, under this condition the uniform Gaussian and the lag-independent priors behave similarly. On the other hand, the ARD prior models the sparsity structure very accurately by assigning a single group to each coefficient. Even though, theoretically it should be able to always properly model the real sparsity of the coefficients, in practice it is beneficial only in the case of very sparse interactions. The drawback of the high complexity of the ARD prior is clearly shown in Figures 5.2 and 5.3 where it appears that ARD overfits the training data. Summarizing, since the lag-independent prior is formulated in order to hold the assumption of the MAR model thus in agreement with the simulated dataset, the fact that it overcomes the other priors is evidence of the effectiveness of GMEP.

GMEP on the empirical dataset Looking at the experiments on the empirical data, i.e. Figure 5.4, the lag-independent prior performs consistently better under different data lengths than the other two priors. This result is in agreement with the outcomes of the comparison on the simulated datasets. And it suggests that the assumption of time independence of the causal configuration, is more plausible than assuming a shared or completely independent sparsity structure. Moreover, the improvement given by the inclusion of the hemisphere partitioning in the structured prior, confirms our assumption that the sparsity structure of the coefficients reflects the hemisphere structure.

Supervised parametric approach for causal inference We developed a classification-based method by assuming a model for the stochastic process and a causality measure for the mapping in the feature space. The idea is to generate a representative dataset of the actual context of which we want to infer the causal interactions and then to map this dataset in the predefined feature space. After that, a classifier is trained in order to predict the causal graph of a given set of time series. This implies that the inference is directly dependent both on the chosen generative model and on the features of the mapping.

Simulated analyses of the supervised method We put this general framework into context by customizing it in the case of the Geweke causal inference in time. This implies the choice of the autoregressive model as generative process of multivariate time series and the assumption of precedence and predictability in time for the identification of a causal interaction.

As Table 5.1 shows, GCA is more sensitive to the additive noise than the supervised approaches. Moreover, Figure 5.5 confirms that the supervised methods (CBC and MBC) perform better than GCA, indeed their curves are closer to the optimal curve than the curve of GCA. And in particular, the inference done on $\mathbf{L}$ by GCA does not provide an estimate of its ROC curve from the origin. It emerges that under a certain value of false positive (or true positive) rate it is not possible to decrease by running GCA on this specific dataset. This is because a large amount of interactions are equally ranked and more precisely, they are assessed to be causal interactions with probability one. Indeed in general, GCA tends to overestimate the causal interactions.

Focusing on the results of CBC and MBC, we notice a similar performance among them, even if CBC performs slightly better than MBC. We just remind that the feature space is the same for both methods, the difference is the number of classifiers and classes.

Comments on the feature space The feature space is a crucial aspect of the supervised approach. Here, we focus on how it treats the multivariate nature of the time series.

Differently from the Geweke measure that is a conditioned pairwise method, in the supervised case the multivariate dependencies among time series are encoded in the feature space by evaluating all the possible causality scenarios. Indeed, the evaluated causality scenarios do not only include all the pair combinations of time series but all the pair combinations of causes and effect, in which as a cause there can be from 1 to M time series.

A better insight on this aspect is provided by Figure 5.6 in which the role of the feature space is investigated. Together with the complete feature space that is defined by the causality scenarios in Table 4.1, two reduced versions (pw and c-pw) are considered. As expected, c-pw and pw feature spaces are less accurate than the complete one in detecting the causal graph. And also their order with respect to the complete case, is in compliance with our expectations, i.e. c-pw provides a richer description of the causal graph than pw.

Toward the real application case By definition in the parametric approaches for causal inference, a realistic model of the generative process has to be defined. Considering the previous analyses, this issue was not taken into account because we evaluated the supervised methods by a cross-validation procedure. Thus, the same dataset was used for both the training and testing phases since our focus was mainly on the feature space and on the classification schema. It is anyway important to observe that in the real-case scenario, there will definitely be a bias due to the model uncertainty. To have an insight

on what would happen in the case of a mismatch between models, we firstly considered an example in which the level of noise is the only source of bias between training and validation data.

Looking at the application of CBC to the **L** dataset with the training done on $\mathbf{L}_{\mathrm{MAR}}$, its AUC score drops from 0.92 to 0.85. By the artificial introduction of the model bias, a decrease was expected, but it is interesting to notice that this result is still higher than 0.72 that is the AUC of GCA on **L**.

Example of real application Regarding the application of the supervised method to the neural recordings, from Table 5.2 we see how the AUC score changes according to the sampling frequency and the time window width, i.e. the model order. We notice that in general, the scores are good. In particular, AUCs are higher when the time window width is of 5 and 10 ms, with the exception of the case of the sampling frequency equals to 600 Hz in which the scores are not influenced by the window width. Moreover, we notice that the AUC decreases with both the sampling frequency and the model order. This can be a sign of overfitting the neural signals since a higher number of time points is used in the mapping to the feature space when we move to the bottom right corner of Table 5.2.

It needs to be remembered that this result is based on the validity of the causal chain EC3→CA1→EC5 that is supposed to represent the actual causal configuration matrix of each trial. Moreover, differently from the previous results, now a further bias might be introduced due to how a causal interaction is decoded in the feature space, beyond the violation of the modelling assumption itself.

Scalability issues on the supervised methods We discuss now some practical implications related to the supervised framework in which we have placed the problem of causal inference.

MBC is strongly affected by the number of time series that composes a trial. Indeed, the number of classes is a power of 2 and only the example of $M = 3$ is computationally tractable. This problem related to the scalability of the method is partially solved with CBC. CBC handles the issue of having an exponentially growing number of classes by focusing the classification at the level of the single cell of the configuration matrix. Since each cell is a binary variable, CBC approaches the inference as a binary classification problem. Consequently, the number of classes is constant and the number of classifiers has a polynomial growth rate in M. Additionally, thinking on the scalability issue from the side of the feature space, the number of scenarios is again exponential with M, but it is not as problematic as the number of classes was in MBC.

A new neurologically plausible model The last activity that is presented in this thesis, aims to investigate the effect of changing the generative model when causality is inferred by a parametric method. Following the same trace of the previous experiments, as parametric method we refer to the autoregressive implementation of the Granger criterion.

The novelty is on the generative model assumed for the validation dataset. Instead of maintaining the consistency between the data generation and the working hypothesis of the methods, a new generative model is adopted which is not strictly based on the MAR implementation but it is more neuro-physiologically plausible.

Causality was inferred from this dataset by GCA and CBC (CBC was initially trained on the **L** dataset). As we can see from Figure 5.7 and Table 5.3 GCA[**NN**] and CBC[**L** $\rightarrow$ **NN**] are very similar in terms of performance level. This can be explained by considering that both methods assume as stochastic model the MAR model and also the inference phase is based on the MAR implementation, even though they derive from two different approaches.

Effect of a changing the generative model The next experiment takes advantage from the possibility offered by CBC to disentangle the generative model from the criterion used to decode a causal interaction. While in the unsupervised approach this is not allowed since the causal criterion is directly derived from the generative model, in the case of the supervised one the training phase allows the chosen causal criterion to be shaped on the adopted generative model.

By keeping the same feature space, i.e. the same causal criterion to identify a connection, CBC was trained on a dataset generated by the **NN** model. The advantage of using the same generative model both in the training and evaluation phases is shown in Figure 5.7 and Table 5.3 where this experiment is labeled as CBC[**NN**].

This improvement of the inference capability was actually expected, since we reduced the bias between training and evaluation phases. Beyond the fact that the effect of this bias reduction is now quantified, it is important to remark the neuro-physiological plausibility of the model. Thinking on the real application scenario, this outcome provides evidence that the supervised approach associated with a neural population model may improve the inference without the need to invert or identify the neural model.

This represents just an initial step before reaching the stage where the causal inference is correctly performed by combining the supervised approach and the neural model, but important for moving in this direction.

6.2 Conclusion

This thesis focused on the problem of time series causality. Specifically, its purpose was to develop new methods for the inference of causal interactions between neural recordings.

More in details, we focused on a well-establish criterion of causality, the Granger criterion, and we considered some issues in its implementation and related new approaches.

In the first part (Part I) of the manuscript, we presented an overview of the research activities that we carried out. Many details were omitted since the aim is to give an high-level description of the work. For a complete description of the methods, datasets, experiments and results, refer to Part II that contains the manuscripts of the three activities.

Summarizing, we addressed the problem of validating the effect of the group sparsity prior in GMEP with simulated dataset. And we showed an application with real fMRI data in which different hypotheses concerning the sparsity structure were tested. The aim of this activity was to characterize GMEP so that it can be applied in the context of MAR identification as initial step of a Granger-based estimate of the causal brain connectivity.

Moreover, we developed a new supervised method for causal graph estimation. It was studied under the Granger definition of causality and compared with the standard (unsupervised) Geweke measure. Two variations were proposed of the supervised method: CBC and MBC. The experiments that we run on these methods, aimed to analyze the roles of the feature space, the classification schema, i.e. MBC vs. CBC, and the generative model of the representative dataset.

The third analysis that concludes this dissertation, aimed to evaluate the effect of adopting a neuro-physiologically plausible model to generate the representative dataset. Firstly, we evaluated the inference performance of both approaches, i.e. unsupervised and supervised, when strictly based on the Granger criterion and applied on neural network data. Then, we exploited the possibility offered by the supervised approach to separate the generative model from the criterion used to decode a causal interaction. Thus, we used the neural network model also for the generation of the representative dataset while the same Granger-based feature space was kept. We believe in the importance of this experiment even though it only involved simulated data, since it goes in the direction of making the inference more interpretable from the neuroscience point of view. And thus it is closer to the principles of effective connectivity.

6.3 Future works

Possible extensions of this research and future works.

- The analysis on GMEP presented here focuses on the effect of changing the group sparsity prior. However, in order to perform a causal connectivity analysis, a further step is needed for computing the causal configuration matrix from the GMEP estimates. In Chapter 7 in which the complete work on GMEP is presented, a solution to this step is proposed. It is based on the idea of considering the posterior distribution of each group, and deriving the presence of an interaction according to its shape. This approach is limited to the use of the lag-independent prior and it was defined by a heuristic evaluation of the simulated results. Thus, it deserves further investigations to evaluate its applicability on real data.

- Regarding the supervised approach for causal inference, the current experiments always consider a trial as a set of 3 time series. An interesting direction for future studies is the extension to a larger number of time series. As we said in Section 6.1, CBC partially handles the problem of having an exponentially growing number of classes. But the generation of a proper representative dataset and the effect that a higher number of time series could have on the feature space, still remain open questions. We already evaluated the idea of tackling the scalability issue by decomposing the causal inference among n time series as $\binom{n}{3}$ inferences among 3 time series. From this side, preliminary results have been collected.

- Beyond the characterization with simulated datasets of the supervised approach, its evaluation with real data is of a great importance. This aspect needs to be analyzed in particular when the neuro-physiologically plausible model is used to generate the representative dataset.

Part II

Papers

Chapter 7

Bayesian Estimation of Directed Functional Coupling from Neural Time Series

Bayesian Estimation of Directed Functional Coupling from Neural Time Series

Danilo Benozzo, Pasi Jylänki, Emanuele Olivetti, Paolo Avesani, Marcel A. J. van Gerven

***Abstract*—In many fields of science, there is the need of assessing the causal influences among time series. Especially in neuroscience, understanding the causal interactions between brain regions is of primary importance. A family of measures have been developed from the parametric implementation of the Granger criteria of causality based on the linear autoregressive modelling of the signals. *Objective:* we propose a new Bayesian method for linear model identification with a structured prior (GMEP) aiming to apply it as linear regression method in the context of the parametric Granger causal inference. *Methods:* GMEP assumes a Gaussian scale mixture distribution for the group sparsity prior and it enables flexible definition of the coefficient groups. Approximate posterior inference is achieved using Expectation Propagation for both the linear coefficients and the hyperparameters. *Results:* GMEP is investigated both on simulated data and on empirical fMRI data. Firstly, GMEP is compared with others standard linear regression methods. And secondly, the causal inferences derived both from GMEP estimates and a standard Granger method, are compared across simulated datasets of different dimensionality, density connection and level of noise. *Conclusions:* these analyses show how adding information on the sparsity structure of the coefficients positively improves the inference process. *Significance:* GMEP allows a better model identification and a consequent causal inference when prior knowledge on the sparsity structure are integrated in the structured prior.**

***Index Terms*—Directed functional connectivity, Granger causality, Bayesian linear regression model, Group sparsity.**

I. Introduction

Wiener-Granger causality is a well-established approach to study causality between time series [1]. This approach is based on the definition of causality proposed by Wiener [2] which considers one time series the cause of another if the latter is better predicted by including information about the first. An implementation of this concept was proposed by Granger [3] who used it to estimate causality between stochastic processes, modelling them as linear autoregressive (AR) models. Specifically, the parametric implementation of Granger causality (GC) identifies a causal interaction between two time series by first modelling them through an AR model and then by comparing how the prediction error changes if each time series is modelled just using its own past values or also including the past values of the others.

Granger causality has been applied in many different fields [4]–[7] and it has become a popular method for identifying causal interactions due to its simplicity and intuitive meaning. This holds particularly in neuroscience, where the understanding of causal interactions among brain areas is of primary importance. According to the terminology adopted in neuroscience, the Wiener-Granger method belongs to the group of the directed functional connectivity methods [8] since it aims to identify the direction of the statistical dependences among a set of brain signals, without making any assumptions about the mechanistic nature of these connections. In neuroscientific applications, given the concurrent acquisition of time series from different brain regions, the problem of inferring causal interactions should take into account the multivariate nature of the data. This desideratum was considered in [9], [10], where a generalization of GC was proposed that relies on the multivariate autoregressive (MAR) model, thereby moving beyond pairwise causal interactions. Apart from the well-known Granger method, several others solutions have been developed. The vast majority of them still starts from the Wiener idea of causality and from modelling the causal interactions through an MAR model [11], [12].

All the approaches that involve MAR modelling require the estimation of the model coefficients as well as of the residual covariance matrix. Since each time point is modelled through a multivariate linear model, this estimation procedure can be shown to be equivalent to solving a multivariate linear regression problem. Due to the nature of neuroscientific datasets, the number of coefficients can be massive. This occurs because signals are acquired from a large number of brain areas. These areas are expressed in term of single or groups of voxels in the fMRI case, and sensors or sources in the MEG/EEG case. Defining d_y as the number of time series and p the so-called order of the MAR model that indicates how many time lags are involved in the modelling of the present time point, the total number of MAR coefficients is $p \times d_y \times d_y$. Hence, the number of unknown coefficients is more than quadratic with respect to the number of time series. This point reveals a crucial property of the multivariate linear regression problem since it is a bottleneck for the scalability of most of the standard linear regression techniques.

The simplest linear regression method is the ordinary least squares (OLS) method. OLS computes the solution by minimizing the root mean square error. There are many examples in which this approach, or variations of it, are considered in the literature [1], [11], [13]–[15]. As mentioned in [16], overfitting is the main risk of OLS when a large number of independent variables are used in the modelling. Even more problematic is the regression if the number of independent variables exceeds the number of observations since the least squares solution will not be unique. Moreover, the high correlation

D. Benozzo, E. Olivetti and P. Avesani are with the NeuroInformatics Laboratory (NILab), Bruno Kessler Foundation, Trento, Italy and with the Center for Mind and Brain Sciences (CIMeC), University of Trento, Italy.

P. Jylänki and M.A.J. van Gerven are with the Radboud University, Donders Institute for Brain, Cognition and Behaviour, Nijmegen, the Netherlands.

between neural time series provides an additional challenge to OLS estimators [17]. In order to overcome the limitations of OLS, one may attempt to regularize the solution [18]. Regularization is done by including in the argument of the cost function a term that controls the overall amplitude of the estimates. This term is generally called penalty term and the resulting approach, penalized regression model. In [19] the authors analysed the use of different penalized regression models, including the well known ridge regression and lasso, for directed functional connectivity estimation. In [18] the elastic net regularizer was considered. Elastic net considers both the penalty terms of ridge and lasso, thus both l_1 and l_2 norms of the coefficients are linearly combined in the cost function. A more sophisticated version of the standard penalized regression models, named group lasso, was proposed in [20] where the authors introduced the concept of grouped variables. By grouped variables, it is meant that the independent variables are clustered in order to find important explanatory variables in predicting the dependent variable. The clustering of the independent variables implies a related clustering of the coefficients at which a separate penalty term is associated. This allows each cluster of coefficients to be separately regularized instead of a global regularization of the whole coefficient vector. This family of methods is often referred to as group sparse regularisation methods. In [21], the asymptotic properties of group lasso were analysed in terms of consistency, normality and uniqueness of the estimate. While in [22], a comparison between standard lasso and group lasso is presented by focusing on the conditions under which group lasso outperforms lasso. In the context of causal inference in multivariate time series, group lasso was studied and compared with non-grouped penalized regression models in [23]. In that work, group lasso was used to enforce coefficient sparsity by grouping together the coefficients connecting the same pair of signals across all time lags. An example application of group lasso with pseudo-EEG data is discussed in [16].

In the Bayesian setting, regularization can be interpreted as imposing a particular prior on the model coefficients. As pointed out in [24], major advantages of Bayesian inference are: the possibility to include prior knowledge in the model definition, the use of model evidence as a measure to compare hypotheses, and finally a quantification of residual uncertainty as captured by the posterior distribution. Several Bayesian approaches were presented in the literature for group sparse modelling, in which the idea of structured priors is exploited to enforce sparsity on the coefficients. The concept of structured (or group sparsity or sparsity-enforcing) priors in the Bayesian setting conveys the same idea of grouped variables. Thus, a structured prior refers to a clustering of the coefficients in which elements in the same group are drawn from the same prior distribution. In [25] a multivariate Gaussian prior was assumed for each group and the expectation maximization (EM) algorithm was used for the inference. A similar approach is presented in [26] where a Dirichlet process prior was employed as structured prior while variational Bayesian was used for the estimate. Other examples have been developed in [27]–[29]. Regarding the application of group sparsity promoting methods in the context of neuroscience, we mention the approach proposed in [30]. In that case, a multidimensional Gaussian distribution was associated to the structured prior and the inference was done in the variational Bayesian framework. This approach was also used in [31]. Another example of a sparse Bayesian regression method is that of [17]. Here, the authors assume that the coefficients are spatially smooth within each time lag and a closed-form solution is obtained by using conjugate priors. The spike-and-slab distribution represents yet another way to constrain the amplitude of the coefficients. This distribution is investigated in [32], [33] as sparsity-enforcing prior for linear regression.

Here, we propose a novel approach for Bayesian group sparse modelling, called GMEP [1]. The name GMEP refers to the Gaussian scale Mixture distribution that is adopted to form a general class of group sparsity priors, and to the Expectation Propagation framework that is used as an efficient method for approximate Bayesian inference. The model is formulated in a general way that enables flexible definition of various non-conjugate observation models. Furthermore, structured priors can be specified using hyperparameters that themselves rely on a multivariate Gaussian prior. The hierarchical structure of the model allows the priors and the hyperparameter vector not to be fixed but modelled by the chosen prior distributions. The posterior is approximated using EP [34] for both the linear coefficients and the hyperparameters. EP has shown to be very accurate and reasonably fast with respect to variational Bayes and Markov chian Monte Carlo [35]. A drawback of EP is the no guarantee of convergence but if properly implemented, convergence can be reliably reached [36].

In this paper, we use GMEP as the basis for a linear regression model to identify a MAR model and to infer the connectivity structure of a given sample of time series. The resulting approach is evaluated both on simulated and empirical fMRI data. The analysis on the simulated dataset aims firstly to compare GMEP with the most commonly used linear regression methods for MAR estimation. Then our approach is evaluated under different prior definitions that represent different sparsity structures of the coefficients. Moreover, we compare the predictive capability of GMEP, and of the multivariate Granger Causality toolbox (MVGC) [15], across different noise levels. Finally, the experiments conducted on the empirical fMRI dataset are meant to investigate the plausibility of some hypotheses related to the sparsity structure of the MAR coefficients. The most realistic hypothesis among the considered ones, is chosen to estimate the directed functional structure in the fMRI time series.

II. Methods

In this section we present the multivariate autoregressive model (MAR) that was used to generate the simulated datasets. Next, a description of the Gaussian Mixture Expectation Propagation (GMEP) method is provided.

A. Multivariate autoregressive model

Let $\mathbf{y}_t$ denote a $d_y \times 1$ vector, representing the state of d_y time series measured at time t. A MAR model of order p,

[1] https://github.com/ccnlab/GMEP

computes $\mathbf{y}_t$ as the linear combination of its p previous time points:

$$\mathbf{y}_t = \sum_{i=1}^{p} \mathbf{A}_i^{\mathrm{T}} \mathbf{y}_{t-i} + \mathbf{e}_t \,, \tag{1}$$

where $\mathbf{e}_t \sim \mathcal{N}(\mathbf{0}, \mathrm{diag}(\sigma_1^2, ..., \sigma_{d_y}^2))$ is the so-called innovation process, its increments are temporally independent and each time instant is a realization from a d_y-dimensional Gaussian distribution with zero mean and diagonal covariance matrix. The $\mathbf{A}_i \in \mathbb{R}^{d_y \times d_y}$ with $i = 1, 2, \ldots, p$ are the coefficient matrices that model the influence of the signal values at time $t - i$ on the current signal values at time t. Thus each $\mathbf{A}_i$ is involved in the data generating process associated with time lag i.

The so-called standard form of the model can be easily derived by constructing the $(d_y p) \times 1$ vector $\mathbf{x}_t = [\mathbf{y}_{t-1}^{\mathrm{T}}, \mathbf{y}_{t-2}^{\mathrm{T}} \ldots \mathbf{y}_{t-p}^{\mathrm{T}}]^{\mathrm{T}}$. $\mathbf{x}_t$ contains the past dynamics of each time series needed to compute the current amplitude $\mathbf{y}_t$. All the $\mathbf{A}_i$ coefficient matrices of each time lag are vertically stacked in a unique $(d_y p) \times d_y$ matrix $\mathbf{W} = [\mathbf{A}_1; \ldots; \mathbf{A}_p]$. Thus, each $\mathbf{y}_t$ is equal to

$$\mathbf{y}_t = \mathbf{W}^{\mathrm{T}} \mathbf{x}_t + \mathbf{e}_t \,, \tag{2}$$

which shows that the model can be identified by solving a multivariate linear regression problem.

B. Gaussian scale Mixture Expectation Propagation method

We present a novel expectation propagation approach for sparse hierarchical generalized linear models and use it as a linear regression method for MAR model identification. Our approach was originally implemented in a more general way that allows the definition of various observation models and coefficient priors. Here, a summary of the method is presented in a context suitable for MAR modeling with a Gaussian observation model and a Gaussian scale mixture distribution for the group-sparsity prior. We will refer to it as GMEP. A detailed description of the model in its general form is given in the Supplementary Material.

As shown in (2), for MAR modeling purposes it suffices to consider a linear regression problem with multiple output variables, where the probability density of each observed $d_y \times 1$ output vector $\mathbf{y}_i$ depends on the $d_x \times 1$ input vector $\mathbf{x}_i$ through a linear transformation $\mathbf{W}^{\mathrm{T}} \mathbf{x}_i$, and $\mathbf{W}$ is a $d_x \times d_y$ matrix of unknown coefficients. We assume that the observation noise is Gaussian and independent over different output variables as well as observations. Therefore, given n input-output pairs, denoted by $\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n$, the observation model can be written as

$$\begin{aligned} p(\mathbf{Y}|\mathbf{X}\mathbf{W}, \boldsymbol{\theta}) &= \prod_{i=1}^{n} p(\mathbf{y}_i|\mathbf{W}^{\mathrm{T}}\mathbf{x}_i, \boldsymbol{\theta}) \\ &= \prod_{i=1}^{n} \prod_{k=1}^{d_y} \mathcal{N}\big(y_{i,k}|\mathbf{w}_k^{\mathrm{T}}\mathbf{x}_i, \underbrace{\exp(\mathbf{V}_{j(i,k)}^{\mathrm{T}}\boldsymbol{\theta})}_{=\sigma_k^2}\big), \end{aligned} \tag{3}$$

where $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_n]^{\mathrm{T}}$ is a $n \times d_y$ output variable matrix, $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_n]^{\mathrm{T}}$ is a $n \times d_x$ input variable (or design matrix) matrix, and $\mathbf{W} = [\mathbf{w}_1, ..., \mathbf{w}_{d_y}]$ is a $d_x \times d_y$ coefficient matrix. In the case of a MAR model, index i enumerates all observed time instants up to n, and d_y corresponds to the number of interacting signals. We assume that each of the nd_y likelihood terms depends on the hyperparameters $\boldsymbol{\theta}$ via a linear transformation by a known $d_\theta \times 1$ vector $\mathbf{V}_{j(i,k)}$ where $j(i,k) = (i-1)d_y + k$, and that the noise level for each output is encoded as $\sigma_k^2 = \exp(\mathbf{V}_{j(i,k)}^{\mathrm{T}}\boldsymbol{\theta})$. Here we simply assume that the noise level can differ between signals but that the noise variance is constant over time points. This can be achieved by including one noise parameter for each output in $\boldsymbol{\theta}$ and by making $\mathbf{V}_{j(i,k)}$ a binary vector that picks the desired component from it for each likelihood term.

The hierarchical prior distributions is of the form $p(\mathbf{W}|\boldsymbol{\theta}) \propto \prod_{j=n+1}^{n+m} p(\mathbf{U}_j^{\mathrm{T}}\mathbf{w}|\mathbf{V}_j^{\mathrm{T}}\boldsymbol{\theta})$, where $\mathbf{w} = \mathrm{vec}(\mathbf{W})$ is a $d_w \times 1$ coefficient vector obtained by vertically concatenating the columns of $\mathbf{W}$. The known transformation matrices $\mathbf{U}_j$ and $\mathbf{V}_j$ are assumed to yield low-dimensional scalar random variables suitable for efficient inference using EP. For MAR identification we adopt a structured Gaussian scale-mixture prior of the form

$$p(\mathbf{W}|\boldsymbol{\theta}) = \prod_{k=1}^{d_y} \prod_{l=1}^{d_x} \mathcal{N}\Big(w_{l,k}|0, \exp\left(\mathbf{V}_{j(l,k)}^{\mathrm{T}}\boldsymbol{\theta}\right)\Big), \tag{4}$$

where $j(l,k) = n + (k-1)d_x + l$ and the prior variance of coefficient $w_{l,k}$ is controlled by $\exp(\mathbf{V}_{j(l,k)}^{\mathrm{T}}\boldsymbol{\theta})$. In GMEP this is obtained by setting $\mathbf{U}_j$ to be unit vectors that pick only one coefficient at a time and $\mathbf{V}_j$ to be binary indicator vectors that cluster the coefficients into a certain number n_g of predefined groups. Each of the groups is assigned an unknown variance hyperparameter $\exp(\theta_{g(j)})$ that is picked up by the inner product $\theta_{g(j)} = \mathbf{V}_j^{\mathrm{T}}\boldsymbol{\theta}$ for each coefficient.

We assign a fixed multivariate Gaussian prior density to the hyperparameters $\boldsymbol{\theta}$:

$$p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}_{\theta,0}, \boldsymbol{\Sigma}_{\theta,0}), \tag{5}$$

where $\boldsymbol{\mu}_{\theta,0}$ is the hyperprior mean vector and $\boldsymbol{\Sigma}_{\theta,0}$ the hyperprior covariance matrix. By adjusting $\boldsymbol{\mu}_{\theta,0}$ and $\boldsymbol{\Sigma}_{\theta,0}$ we can form coefficient priors with different sparsity-promoting properties. For example, if we set $\mathbf{V}_{j(l,k)}$ to unit vectors that attach only one hyperparameter to each coefficient and assume $\boldsymbol{\Sigma}_{\theta,0}$ to be diagonal, we can create sparser solutions by increasing the diagonal entries of $\boldsymbol{\Sigma}_{\theta,0}$ and decreasing the prior means $\boldsymbol{\mu}_{\theta,0}$. An uninformative signal-specific noise prior can be obtained by making the corresponding elements of $\boldsymbol{\mu}_{\theta,0}$ sufficiently small and including an "independent" diagonal block in $\boldsymbol{\Sigma}_{\theta,0}$ with sufficiently large diagonal values. This corresponds to setting independent log-normal priors to the noise variances σ_k^2.

Figure 1 shows the graphical model representation of GMEP. Random variables are denoted with circles, while known variables are denoted with rectangles. The fixed hyperpameters $\boldsymbol{\mu}_{\theta,0}$ and $\boldsymbol{\Sigma}_{\theta,0}$ are denoted with dots.

This general model definition enables the implementation of various different linear models via the choice of the transformations $\mathbf{V}_1, ..., \mathbf{V}_n$ for the likelihood terms and $\mathbf{V}_{n+1}, ..., \mathbf{V}_{n+d_y d_x}$ for the prior terms. In the following we present the three structured coefficient priors that were used

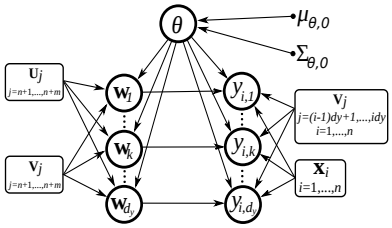


Fig. 1: Graphical model of GMEP in which dependences between variables are shown by using circles for random variables, rectangles for known variables and dots for fixed hyperparameters.

in the experiments. They are described in a formal and mathematical way by considering the notation adopted until now, their actual interpretation and meaning are reported in Subsection III-C.

1) Uniform Gaussian prior: A uniform Gaussian prior with unknown scalar prior variance for each output (similar to ridge regression) can be obtained by choosing $\mathbf{U}_j = \mathbf{e}_j$ $(d_w \times 1)$ and $\mathbf{V}_j = \mathbf{e}_k$ $(d_y \times 1)$ for $j = (k-1)d_x+n+1, ..., (k-1)d_x+n+d_x$ and $k = 1, .., d_y$. This leads to $n_g = d_y$ different inference problems, if the coefficients related to different outputs are not coupled through the observation model;

2) Automatic relevance determination: An automatic relevance determination (ARD) prior can be formed by assigning individual scale hyperparameters to each coefficients. Thus, we set $\mathbf{U}_j = \mathbf{V}_j = \mathbf{e}_j$ $(d_w \times 1)$ with $j = n+1, ..., n+d_w$. This construction assumes individual scale parameters for each coefficient $n_g = d_w$ and no information sharing between the outputs, which results in independent regression problems for each output. This prior is very flexible because each of the $d_w = d_y d_y p$ coefficients can be regularized out of the model independently, but the resulting inference problem is also more challenging in terms of avoiding overfitting.

3) Group sparsity prior: Group sparsity priors can be constructed by defining possibly overlapping groups as $\mathbf{U}_j = \mathbf{e}_j$ $(d_w \times 1)$ and $\mathbf{V}_j = [1, 0, 1, 0, 0, ..., 0]^{\mathrm{T}}$ $(n_g \times 1)$. Groups could be defined either so that they combine coefficients from different output units into same groups or completely separately for each output. In particular, in our experiments we will use a group sparsity prior defined by choosing $\mathbf{U}_j = \mathbf{e}_j$ $(d_w \times 1)$ and $\mathbf{V}_j = \mathbf{e}_{(r-1)d_y+l}$ $(d_y d_y \times 1)$ with $j = (k-1)d_y + (r-1)d_x + l + n$ for $l = 1, ..., d_y$, $r = 1, .., d_y$ and $k = 1, ..., p$. From now on, we will refer to this group sparsity prior as lag-independent sparsity since it assumes that the coefficient sparsity structure is independent from the time lag and also not shared between the outputs. Compared to the ARD prior, the lag-independent sparsity is less flexible because it combines information over different lags. However, it still provides $d_y \times d_y$ free prior parameters that can explain the causality structure between the d_y interacting signals in our MAR model.

C. Approximate inference

A deterministic Gaussian approximation to the posterior distribution is computed using the EP algorithm [34]. The posterior approximation for the GMEP, defined by combining equations (3), (4) and (5), is formed by replacing the non-Gaussian likelihood terms, $\mathcal{N}(y_{i,k}|\mathbf{w}_k^{\mathrm{T}}\mathbf{x}_i, \exp(\mathbf{V}_{j(i,k)}^{\mathrm{T}}\boldsymbol{\theta}))$, and the prior terms, $\mathcal{N}(\mathbf{w}_{l,k}|0,\ \exp\left(\mathbf{V}_{j(l,k)}^{\mathrm{T}}\boldsymbol{\theta}\right))$, with joint Gaussian functions of $\mathbf{w}$ and $\boldsymbol{\theta}$. Note that if $\boldsymbol{\theta}$ was known, no EP approximation would be needed since the terms of the model are already Gaussian with respect to $\mathbf{w}$.

The EP algorithm proceeds by initializing the approximate factors to some sensible values, and then updates each of them in turn. At each update, first, one of the approximate terms is removed from the approximation and replaced with the actual model term to give a tilted distribution, which can be regarded as a more refined approximation to the posterior. Then the parameters of the left-out approximate term are updated so that the KL divergence from the tilted distribution to the true distribution is minimized. In case of a Gaussian approximation this corresponds to matching the mean and covariance of the approximation with the tilted distribution. This iteration is repeated at some order for all model terms until convergence. In practice we update all the likelihood terms in one batch keeping the prior term approximations fixed, and vice versa. Finally, after convergence, posterior summaries of the unknown model parameters and predictions are computed using the Gaussian approximation for $\mathbf{w}$ and $\boldsymbol{\theta}$.

III. Materials

The first two parts of this section describe the simulated and empirical datasets that were used in the experiments. Whereas the last part is about the structured coefficient priors adopted in GMEP.

A. Simulated MAR datasets

The synthetic datasets were generated by an MAR model, and our goal is to study how good is the identification of GMEP. In order to explore the model performance in different regimes, multiple ensembles of time series were generated under different conditions. In our simulations the free parameters that identify a dataset are the dimensionality d_y and the connection density c. Here, d_y refers to the number of time series contained in each trial of the dataset and c refers to the fraction of non-zero off-diagonal connections (i.e. causal interactions). This choice to characterise each dataset through the pair (d_y, c) is motivated by the fact that it heavily influences the ability to accurately estimate causal interactions.

In our simulations $d_y \in \{3, 7, 11\}$ and $c \in \{0.1, 0.5, 0.9\}$. Each dataset, indexed by (d_y, c), consists of 100 trials (repetitions). Each trial $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]^{\mathrm{T}}$ is a $n \times d_y$-dimensional matrix, where the length of each to the d_y time series is set to $n = 1500$ time points. $\mathbf{Y}$ is generated by an MAR model of the predefined order $p = 10$ and with a predefined causal configuration matrix $\mathbf{A}$. $\mathbf{A}$ is a binary matrix, it contains the causal structure that determines the interactions between time series. Specifically, $\mathbf{A}(r, s) = 1$ means that signal r causes signal s. In each time lag, the related $\mathbf{A}_i$ matrix is generated

by multiplying the non-zero elements of $\mathbf{A}$ with Gaussian distributed random numbers.

Each trial has its own configuration matrix $\mathbf{A}$ while the connection density c is shared between trials in the same dataset. Not all the n time points are used in the analyses since we decided to keep the same proportion of elements in the design matrix and unknowns (coefficients) in order to have more comparable results across datasets. Thus the number of actual time points involved in the experiments depends on d_y. In Table I, we report for each d_y the related n and the resulting shape of $\mathbf{Y}$, $\mathbf{X}$ and $\mathbf{W}$.

d_y	n	$\mathbf{Y} : [n \times d_y]$	$\mathbf{X} : [n \times (d_y p)]$	$\mathbf{W} : [(d_y p) \times d_y]$
3	189	189×3	189×30	30×3
7	441	441×7	441×70	70×7
11	693	693×11	693×110	110×11

TABLE I: For each d_y the number of time points n is specified and the resulting shape of matrices $\mathbf{Y}$, $\mathbf{X}$ and $\mathbf{W}$.

In the last experiment conducted on the simulated data we introduced a third free parameter: the level of noise γ, $\gamma \in \{0.2, 0.4, 0.6, 0.8\}$. Each trial is computed as $\mathbf{Y} + \gamma \mathbf{Y_{noise}}$, where $\mathbf{Y_{noise}}$ has the same shape of $\mathbf{Y}$ and it is the output of an univariate AR process.

B. Empirical fMRI dataset

The empirical data we used belong to the *Gallant Lab Natural Movie 4T fMRI Dataset* [37], [38] and were acquired on a 4T Varian INOVA scanner. The scanning was done using T2*-weighted gradient echo EPI: TR=1 s, TE=28 ms, Flip angle=56 degrees, voxel size $= 2.0 \times 2.0 \times 2.5$ mm^3, and FOV=128×128 mm^2. A total of 18 coronal slices were acquired and they cover the posterior portion of occipital cortex, starting at the occipital pole. A parcellation of the measured voxels into 26 regions of interest was provided by the authors. Subjects were presented with natural movies during a training and a test session. See [38] for further details about the experimental protocol.

The time series we used in our analysis were extracted from the training dataset of one of the three acquired subjects by averaging signals corresponding to the same region of interest. This gave for each subject 26 time series; one for each ROI. Each time series had a length of 7200 s, since 12 separate 10-minute blocks of movies were presented for the training dataset. For our analyses, we considered the concatenated block and ignored modelling errors at the boundaries between blocks.

C. Employed structured coefficient priors

In Section II, we explained the structured priors in analytical terms, here we will recall them giving an interpretation of their analytical definition from the point of view of the sparsity structure that they assume.

Firstly, a uniform Gaussian prior was defined for each output. Such configuration is strictly related to ridge regression because the coefficients associated at each output are supposed to belong to the same group that means they are modelled as drawn from the same distribution. This implies that sparsity is shared across all coefficients in the same column of $\mathbf{W}$ since only the hyperparameters that define such distribution tune the level of sparsity. In other words, we can see this as the GMEP implementation of ridge.

The second trivial configuration that was taken into account, considers one group for each coefficient. It represents the opposite situation with respect to the previous prior, thus now each coefficient has its own distribution to which it belongs to. This approach is known as automatic relevance determination (ARD) because the hyperparameters of each distribution determine the sparsity i.e. the relevance, of the related coefficient.

The third case in our comparison has a definition of groups that reproduces the true sparsity structure of the coefficients in $\mathbf{W}$. Referring to Equation 1 and to the description of how each $\mathbf{A}_i$ was computed from $\mathbf{A}$, we can see that the same sparsity structure is shared across time lags, i.e. the amount and position of the zero connections are the same across $\mathbf{A_i}$. This assumption can be rephrased as: the causal configuration is time independent, i.e. there is no dynamic in the causal interactions. Therefore, we call that prior group the lag-independent prior.

IV. Experiments

This section describes the experiments that were run to analyse GMEP and to study its application both on simulated and empirical data.

A. Simulated MAR datasets

We start with the experiments that were run on the synthetic data. As described below, these experiments have three unique purposes.

The first purpose is to compare GMEP and other standard linear regression approaches. In particular, we refer to Ordinary Least Squares (OLS), Levinson-Wiggs-Robinson equations (LWR) and Ridge Regression (RR). They are all standard methods widely used for linear regression. In particular, OLS and LWR are both used in practice to fit the MAR parameters in MVGC. Moreover, both are point estimator methods and asymptotically equivalent to the maximum likelihood estimate. The last technique, RR, is included since it contains a regularization term in order to prevent overfitting.

Note that LWR derives from a multivariate extension to Durbin recursion and it has the advantage to provide also an estimate of the residual covariance matrix $\hat{\mathbf{\Sigma}}$. For further details refer to [15], [39]. RR can be simply expressed by adding the l_2-norm of the coefficient matrix $\mathbf{W}$ in the objective function of OLS [40]. In this way, the magnitude of the coefficient is included in the minimization process and it is forced to be small according to a weight parameter that controls the amount of shrinkage. The comparison of GMEP, OLS, LWR and RR is done by running them on each synthetic dataset and focusing on their capability to estimate the coefficient matrix. The model order is set equal to its true value, i.e. $p = 10$, and the performance of each approach is evaluated through the normalized root mean square error (NRMSE) computed

between the true and the mean of posterior distribution of the estimated coefficients. The normalization is done according to the maximum amplitude (the difference between the maximum and minimum) of the true coefficients. Hence, NRMSE is not necessarily bounded in $[0, 1]$.

The second purpose of the experiments on the simulated data is to focus exclusively on GMEP and analyse the impact of the structured priors. In particular, the aim is to show that a more detailed modelling of the group sparsity prior, through the inclusion of information related to the structure of the data, improves the results. Thus, we are interested in proving that there are situations in which an accurate definition of the structured prior leads to a better inference. This improvement is observable not only through a comparison with the true coefficients but also by evaluating the reconstructed time series. To understand how the structured priors affect the final results, a comparison across three different priors is conducted. We refer to Subsection III-C for details on the structured coefficient priors.

The predictive performances of the different priors are evaluated by computing the mean log predictive densities (MLPD):

$$\frac{1}{n_t d_y}\sum_{i=1}^{n_t}\sum_{j=1}^{d_y}\log\int p(y_{i,j}^*|\mathbf{w}_j^{\mathrm{T}}\mathbf{x}_i^*,\exp(\mathbf{v}_{i,j}^{\mathrm{T}}\boldsymbol{\theta}))p(\mathbf{w},\boldsymbol{\theta}|\mathcal{D})d\mathbf{w}d\boldsymbol{\theta}$$
$$\approx\sum_{i=1}^{n_t}\sum_{j=1}^{d_y}\log\int p(y_{i,j}^*|\mathbf{w}_j^{\mathrm{T}}\mathbf{x}_i^*,\exp(\mathbf{v}_{i,j}^{\mathrm{T}}\boldsymbol{\theta}))q(\mathbf{w}_j)q(\boldsymbol{\theta})d\mathbf{w}_j d\boldsymbol{\theta},$$

where $\mathbf{y}_i^* = [y_{i,1}^*, ..., y_{i,d_y}^*]^{\mathrm{T}}$ is a known test observation at test input $\mathbf{x}_i^*$, and $q(\mathbf{w},\boldsymbol{\theta}) = \prod_j q(\mathbf{w}_j)q(\boldsymbol{\theta})$ is given by the EP approximation. With a Gaussian observation model the required integrals can be computed using one dimensional numerical quadratures. Higher MLPD values correspond to higher approximate predictive density values for test data points on average indicating thus better predictive performance. During the EP iterations we repeatedly evaluate the normalization coefficients $\hat{Z}_{i,j}$ of the tilted distributions, which for likelihood terms are defined as

$$\hat{p}_i(\mathbf{w},\boldsymbol{\theta}) = \hat{Z}_{i,j}^{-1}p(y_{i,j}^*|\mathbf{w}_j^{\mathrm{T}}\mathbf{x}_i^*,\exp(\mathbf{v}_{i,j}^{\mathrm{T}}\boldsymbol{\theta}))q_{-i}(\mathbf{w},\boldsymbol{\theta}),$$

where

$$\hat{Z}_{i,j} = \int p(y_{i,j}^*|\mathbf{w}_j^{\mathrm{T}}\mathbf{x}_i^*,\exp(\mathbf{v}_{i,j}^{\mathrm{T}}\boldsymbol{\theta}))q_{-i}(\mathbf{w},\boldsymbol{\theta})d\mathbf{w}d\boldsymbol{\theta}.$$

Since the cavity distributions $q_{-i}(\mathbf{w},\boldsymbol{\theta})$ can be regarded as an approximation to the posterior when observation $y_{i,j}$ is left out from the training set, we can use the normalisation terms $\hat{Z}_{i,j}$ to form an approximation to the mean leave-one-out predictive densities:

$$\mathrm{MLPD}_{\mathrm{EP}} = \frac{1}{nd_y}\sum_{i=1}^{n}\sum_{j=1}^{d_y}\log\hat{Z}_{i,j}.$$

In the experiments we use $\mathrm{MLPD}_{\mathrm{EP}}$ as an estimate of the future predictive performance of the model and validate it with respect to the actual MLPD score using simulated experiments.

For a known coefficient vector $\mathbf{w}^*$, a similar measure that we call $\mathrm{MLPD}_{\mathbf{w}}$ can be computed as

$$\mathrm{MLPD}_{\mathbf{w}} = \log\int p(\mathbf{w}^*,\boldsymbol{\theta}|\mathcal{D})d\boldsymbol{\theta} \approx \sum_{j=1}^{d_y}\log q(\mathbf{w}_j^*),$$

which measures how well the posterior approximation matches with the true coefficients. A higher $\mathrm{MLPD}_{\mathbf{w}}$ value indicates a better agreement with the EP posterior approximation $q(\mathbf{w})$ and the true coefficients $\mathbf{w}^*$.

The third purpose of the experiments conducted on the simulated data is to obtain an estimate of the binary causal configuration matrix from the results of GMEP. Such analysis requires the choice of a specific structured prior and a way to obtain the binary causal configuration matrix from the results of the inference process. The proper structured prior is chosen according to the outcome of the previous experiment by selecting the one with the best results, as we will see later it is the lag-independent prior. And the binary causal configuration matrix is computed by considering that an estimate of the variance distribution of each group is provided by GMEP. In detail, due to the choice of the lag-independent prior each group contains all the coefficients that link the same pair of time series at different time lags. Thus, there is one group for each cell of the binary configuration matrix. Moreover, the coefficients in each group are supposed to be normally distributed with zero mean and variance that is modelled as a log-normal distribution. After the estimation process, the posterior mean and standard deviation of such distribution are used to reconstruct the causal configuration matrix of each trial by their normalization and comparison. The causal configuration matrices predicted by MVGC and the ones predicted by GMEP are evaluated with the related ground truth. Such comparison is extended also to the datasets with the noise component, thus to all the (d_y, c, γ) datasets. From the estimated causal configuration matrix of each trial, the true positive rate and true negative rate are computed and averaged across trials with the same level of noise. This procedure was repeated both for MVGC and GMEP, in order to compare them in term of their balanced accuracy (BA). We chose the balanced accuracy, as evaluation measure since it overcomes the problem of unbalanced dataset [41]. BA is meant as the mean of the true positive rate and the true negative rate across all the trials in each (d_y, c, γ) dataset.

B. Empirical fMRI dataset

The second part of the experiments focuses on the empirical data. We are aware of the existing debate about using time lag-based method with fMRI data. Indeed a number of studies state that the BOLD response is not compatible with the assumptions of precedence and predictability that are at the root of Granger causality [42], [43], while others prove the robustness of Granger causality to variations of the hemodynamic response function and identifies the noise level and the amount of downsampling as possible issues in causal prediction [44]. Here, we do not enter this debate but we aim to use the Bayesian model as a way to test hypotheses about the sparsity structure. In this way, if prior knowledge is available

on the structure of the data, then it is possible to test it and to compare the results with respect to a baseline case such as the ARD or the uniform Gaussian priors.

When working with empirical data, the main difference with respect to the analysis conducted on the simulated one, is that we do not have the ground truth on the sparsity structure. This lack of information can be replaced by prior knowledge on the data. For example, it is reasonable to assume a difference in the magnitude of the coefficients that connect areas in the same hemisphere respect to the ones that connect areas across hemispheres. Such an assumption can be encoded in a specific group sparsity prior and a comparison with other structured priors can reveal which is the closest to the ground truth. We defined an experiment in which the length of each time series is gradually reduced in order to compare the performances of GMEP under both different structured priors and number of training time points. In detail, the experiment that we have carried out on the empirical dataset, is the following: first, part of the dataset is used to identify the best model order with a grid search approach. Next, we apply GMEP using the three structured priors that we adopted on the simulated data, see Subsection III-C. Moreover, in the definition of the structured prior we also consider the anatomical position associated with each time series. That is, we enrich the three initial priors by adding four new groups in which the coefficients are clustered according to the hemispheres that they link with. The results of these two scenarios, i.e. the three structured priors and their enrichment with anatomical information, are compared with the prior that only models the hemisphere structure. This allow us to identify the most plausible group sparsity prior among the tested ones. This prior is used in the final analysis, where the aim is to compute the causal configuration matrix by using GMEP. The approach used to obtain a binary matrix is the same as the one used for the simulated data based on the comparison between the posterior mean and posterior standard deviation of the group variances.

V. Results

Results are divided according to the dataset from which they were obtained, thus the first part of this section is devoted to the findings from the simulated MAR datasets and the latter to the empirical fMRI dataset.

A. Simulated MAR datasets

Figure 2 shows the results of the NRMSE computed between estimated and true coefficients by OLS, LWR, RR and GMEP. These four methods were applied at each simulated dataset. The figure reports the median and the 25-th and 75-th percentiles computed on the 100 trials of each dataset. In the case of RR the strength of the regularization term was selected through a grid search approach applied on a subset of time points. For GMEP the uniform Gaussian coefficient prior was adopted. The results show that, while the prediction errors of OLS, LWR and RR do not show large differences, the prediction error of GMEP is consistently better. As expected, we observe that NRMSE increases with the connection density. Moreover, the percentiles are very small thus the prediction error is stable across trials in each dataset and for each regression method.

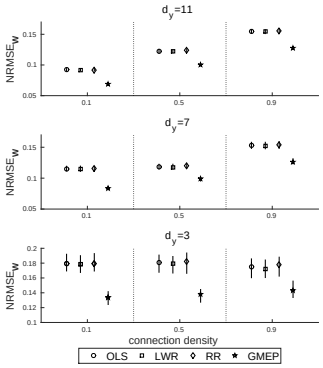


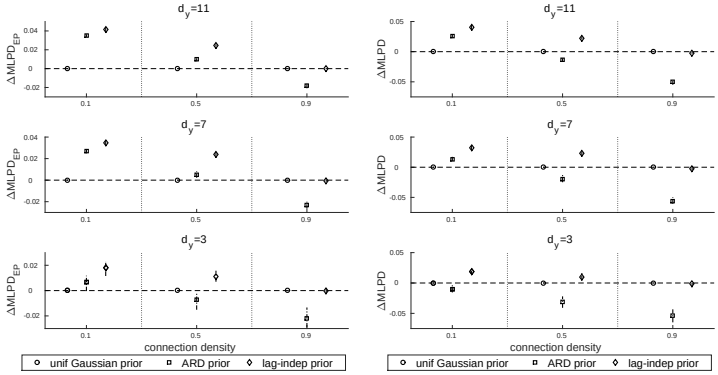
Fig. 2: NRMSE related to the coefficient estimations, each inference method is identified by a specific marker and its result is reported in terms of median, 25-th and 75-th percentiles.

Next, we analysed the performance of GMEP under different coefficient priors. The comparison across priors is done by using the uniform Gaussian prior as a baseline with which other priors are compared. The predictive performance is evaluated through the mean log predictive density (MLPD). In particular, we will consider the variation of MLPD with respect to the uniform Gaussian prior that we indicate as ΔMLPD. In Figure 3 the $\Delta\text{MLPD}_{\mathbf{W}}$ computed on the coefficients is shown. Figure 4 contains both the $\Delta\text{MLPD}_{\text{EP}}$ and the actual ΔMLPD computed on a separated test set.

Figure 3 shows that the ARD prior outperforms the uniform Gaussian prior only for connection density equals to 0.1 and dimensionality equals to 7 and 11. Its performance decreases as connection density is increased. In general, the lag-independent prior performs better than the other priors, particularly for low to medium connection densities. The lag-independent prior becomes comparable to the uniform Gaussian prior in the case of very dense configurations.

The same behaviour is reported in Figure 4. Both the ARD and the lag-independent priors get worse with the increase of the connection density with the difference that the lag-independent prior becomes comparable to the uniform Gaussian prior in the worst case. On the other hand the ARD prior drops faster and only in few cases it is better than the uniform prior. By comparing Figure 4(a) and (b) the generalization capability of the ARD and the lag-independent priors is highlighted. In fact we notice that in the ARD prior the MLPD drops faster than the MLPD_{EP} but in the case of the lag-independent prior MLPD and MLPD_{EP} behave similarly.

Figure 5 shows the difference between the balanced accuracy computed by applying GMEP and MVGC, under different levels of noise. We remember that the balance accuracy BA is defined as the mean of the true positive rate and the true negative rate and we will refer at the difference of BA between GMEP and MVGC as ΔBA. The noise level is quantified by the parameter γ and indicates the proportion between the actual signal and the univariate noise, i.e. $\gamma = 0$ means that



(a) (b)

Fig. 4: ΔMLPD computed with respect to the uniform Gaussian prior and evaluated on the EP iterations and on the test set.

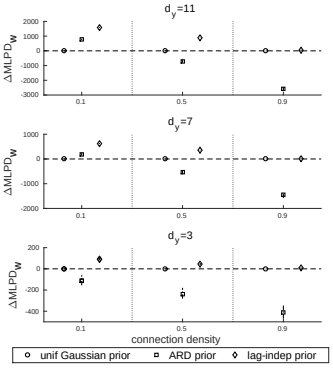


Fig. 3: ΔMLPD computed with respect to the uniform Gaussian prior and evaluated on the coefficient estimates.

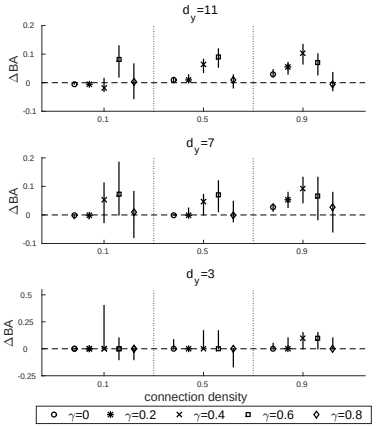


Fig. 5: ΔBA computed on the causal configuration matrices estimated by GMEP and MVGC.

the noise component is absent. As in the previous figures, the median and the 25-th and 75-th percentiles are reported. In this case the marker indicates the level of noise. The figure shows that there are no meaningful differences in the cases of $d_y = 3$, i.e. for low dimensionality. On the other hand, significant differences appear when the noise level increases and in particular when also the connection density increases. At increasing noise levels, the predictions of GMEP become more accurate than the predictions of MVGC. The gap of BA between the two approaches reaches the 10% in favour of GMEP for medium levels of noise ($\gamma = 0.4$ and $\gamma = 0.6$) and it drops to 0 when the data are dominated by the noise, i.e. $\gamma = 0.8$.

B. Empirical fMRI dataset

The experiments conducted on the empirical data, as described in Section IV, are meant to test hypotheses about the sparsity structure of the causal interactions among the brain regions which the analysed time series correspond to. In Figure 6, we report the MLPD under different structured priors and number of training time points. We did not include the equivalent results for other performance measures since they show the same trend. In detail, the first 500 time points were firstly used to determine the order of the MAR model. This analysis showed a good compromise between performance and model complexity for $p = 4$. Using this model order, GMEP was applied in conjunction with the uniform Gaussian prior, the ARD prior and the lag-independent sparsity prior. Figure 6 reports with lines marked by circles the results of these priors using a different colour for each of them. The lines marked by squares show the effect of the inclusion of the partitioning based on the hemispheres. The black line reports the results with only the hemisphere groups in the sparsity structure prior. The results always show an improvement when the hemisphere groups are included in the structured prior. Moreover, consistent with the simulations, the lag-independent prior achieved the highest performance.

Finally, we report the causal configuration matrix that is obtained by the predictions of GMEP. Based on our findings, we adopt the lag-independent prior associated with the hemisphere

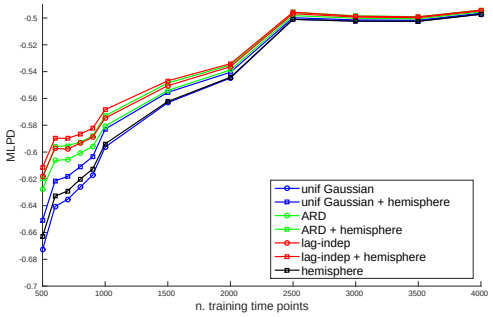


Fig. 6: MLPD on the test set computed by multiple applications of GMEP under different structured priors and by varying the number of time points in the training set.

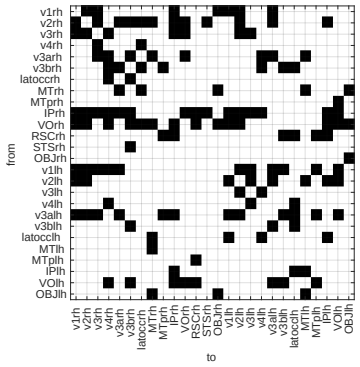


Fig. 7: The causal configuration matrix computed on the empirical fMRI data, the black squares indicate a causal interaction from the rows to the columns.

partition. The configuration matrix is computed by following the same approach that was used in the synthetic data. About the number of time points, the same proportion of elements in the design matrix and unknowns was also preserved for the empirical data. Thus, since in this dataset $d_y = 26$, to be consistent with the previous analyses, 1638 time points were selected for the inference. The causal configuration matrix is shown in Figure 7 and it contains a black square when a causal interaction is determined from a region along the rows to a region along the columns. Based on this matrix, we tested the significance of the sum of the overlapping connections between the two hemispheres, i.e. the intersection of the two sets of connections within hemisphere. And the significance of the sum of the connections of homologous areas across hemispheres. In both cases, the null hypothesis was rejected with a significance level of 0.01. The distribution of the null hypothesis was computed by randomly permuting the estimated connections for 1 million of iterations.

VI. Discussion

In this paper we analysed a novel approach for Bayesian linear modelling with structured prior (GMEP) in the context of the MAR identification with the aim to apply it for a Granger-based estimate of directed functional brain connectivity.

We first made a simple comparison with other standard linear estimators to see how GMEP is placed in relation to them. By evaluating the NRMSE of the coefficient estimates, GMEP showed the most accurate predictions. Our results also provide an insight into how the connection density and the dimensionality influence the inferences. In particular, we obtained that given a certain dimensionality, the complexity of the estimation problem increases with the increase in connection density. It is important to highlight that dimensionality and number of unknowns (coefficients) are related, thus in all of our experiments the proportion between number of elements in the design matrix and unknowns was kept constant across datasets.

One of the main advantages of GMEP is its flexibility in the definition of the structured prior. Thus this aspect was studied through several simulations. The simulations were meant to test how the structured prior affected the predictions under different conditions of dimensionality and connection density. In the case of sparse datasets, i.e. datasets with low connection density, modelling the sparsity improves the performance of GMEP.

We modelled the sparsity by two types of structured priors. That is, the ARD prior and the lag-independent prior, which were compared with the uniform Gaussian prior.

The uniform Gaussian and the ARD priors can be seen as two extreme scenarios in terms of model complexity. In the case of the uniform Gaussian prior, the model complexity is very low since all the coefficients that are involved in the modelling of the same time series are clustered in the same group. Thus they are supposed to be drawn from the same distribution, i.e. they are assumed to have the same sparsity. This assumption is realistic only in case of very high connection density. Indeed, under this condition the uniform Gaussian and the lag-independent priors behave similarly. On the other hand, the ARD prior models the sparsity structure very accurately by assigning a single group to each coefficient. Even though, theoretically it should be able to properly model the real sparsity of the coefficients, in practice it is beneficial only in case of very sparse interactions. The drawback of the high complexity of the ARD prior is clearly shown in the Figure 4 where it appears that ARD overfits the training data.

The lag-independent prior was shown to always outperform the other priors or, in the worst case, be equal to the uniform Gaussian prior. This result was expected since such a prior models the actual sparsity structure of the coefficients, forming an optimal compromise in term of model complexity. Summarizing, these results provide evidence of the importance of adding prior knowledge about the sparsity structure of the coefficients in the model.

Regarding the ability to predict the causal interactions among time series, we can conclude that GMEP reaches a balanced accuracy that is the 10% higher than the one of MVGC for some levels of noise. This result is important for the application in empirical settings in which we do not know neither the true amount of noise nor the true connection density. Even though the experiment was restricted to just three

dimensions and a fixed number of time points, it shows that GMEP can provide meaningful advantages, particularly for medium noise levels.

The experiments on the empirical data under the three structured priors showed that, in agreement with the simulations, the lag-independent prior performs consistently better under different data lengths.This evidence suggests that the assumption of time independence of the causal configuration, is more plausible than assuming a shared or completely independent sparsity structure. Moreover, the improvement given by the inclusion of the hemisphere partitioning in the structured prior, confirms our assumption that the sparsity structure of the coefficients reflects the hemisphere structure. Regarding the causal configuration matrix, the simple statistical tests that were run on it, suggest significant symmetries on the connections within and across hemispheres.

VII. Conclusion

A new Bayesian method for linear regression with structured prior (GMEP) was proposed and applied in the context of the MAR identification. The purpose was to identify the MAR model in order to obtain a Granger-based estimate of the causal configuration matrix from a given set of time series. The main advantage of GMEP is a flexible definition of various structured priors associated with the sparsity structure of the MAR coefficients. We investigated GMEP among standard linear estimators on simulated datasets with different dimensionalities and connection densities. Moreover, we focused on the effect of defining different structured priors. And we showed the benefit of including information on the sparsity structure of the coefficients in their prior definition. In the same simulation framework, we identified under with conditions the inference of the causal configuration matrices performed by GMEP achieves better results than the inference done by a standard Granger toolbox (MVGC). Finally, we reported a simple example with empirical fMRI data showing that the enrichment of the structured prior by the inclusion of anatomical information i.e. the hemisphere partitioning, leads to a better inference.

Acknowledgements

This research was supported by grant numbers 612.001.211 and 639.072.513 of The Netherlands Organization for Scientific Research (NWO) and by funds from the Bruno Kessler Foundation (FBK) and the Finnish Cultural Foundation. Moreover, we thank the CRCNS founding program for having made available the fMRI dataset [37], [38].

References

[1] S. L. Bressler and A. K. Seth, "Wiener-Granger causality: a well established methodology." *NeuroImage*, vol. 58, no. 2, pp. 323–329, Sep. 2011. [Online]. Available: http://dx.doi.org/10.1016/j.neuroimage.2010.02.059

[2] N. Wiener, "The theory of prediction," in *Modern mathematics for engineers, Series I*, E. F. Beckenham, Ed. McGraw-Hill, 1956.

[3] C. W. J. Granger, "Investigating Causal Relations by Econometric Models and Cross-spectral Methods," *Econometrica*, vol. 37, no. 3, pp. 424–438, Aug. 1969. [Online]. Available: http://dx.doi.org/10.2307/1912791

[4] R. Nagarajan and M. Upreti, "Comment on causality and pathway search in microarray time series experiment." *Bioinformatics (Oxford, England)*, vol. 24, no. 7, pp. 1029–1032, Apr. 2008. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btm586

[5] A. Rao, A. O. Hero, D. J. States, and J. D. Engel, "Inference of Biologically Relevant Gene Influence Networks Using the Directed Information Criterion," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 2. IEEE, May 2006, p. II. [Online]. Available: http://dx.doi.org/10.1109/icassp.2006.1660521

[6] R. K. Kaufmann and D. I. Stern, "Evidence for human influence on climate from hemispheric temperature relations," *Nature*, vol. 388, no. 6637, pp. 39–44, Jul. 1997. [Online]. Available: http://dx.doi.org/10.1038/40332

[7] C. Sims, "Money, income, and causality," *American Economic Review*, vol. 62, no. 4, pp. 540–52, 1972. [Online]. Available: http://EconPapers.repec.org/RePEc:aea:aecrev:v:62:y:1972:i:4:p:540-52

[8] K. Friston, R. Moran, and A. K. Seth, "Analysing connectivity with Granger causality and dynamic causal modelling." *Current opinion in neurobiology*, vol. 23, no. 2, pp. 172–178, Apr. 2013. [Online]. Available: http://dx.doi.org/10.1016/j.conb.2012.11.010

[9] 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, Jun. 1982. [Online]. Available: http://dx.doi.org/10.1080/01621459.1982.10477803

[10] A. B. Barrett, L. Barnett, and A. K. Seth, "Multivariate Granger Causality and Generalized Variance," *Physical Review E*, vol. 81, no. 4, pp. 041 907+, Apr. 2010. [Online]. Available: http://dx.doi.org/10.1103/physreve.81.041907

[11] S. Haufe, V. V. Nikulin, K.-R. R. Müller, and G. Nolte, "A critical assessment of connectivity measures for EEG data: a simulation study." *NeuroImage*, vol. 64, pp. 120–133, Jan. 2013. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/23006806

[12] 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, Jun. 2001. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/11417058

[13] A. K. Seth, "A MATLAB toolbox for Granger causal connectivity analysis," *Journal of Neuroscience Methods*, vol. 186, no. 2, pp. 262–273, Feb. 2010. [Online]. Available: http://dx.doi.org/10.1016/j.jneumeth.2009.11.020

[14] A. Schlögl, "A comparison of multivariate autoregressive estimators," *Signal Processing*, vol. 86, no. 9, pp. 2426–2429, Sep. 2006. [Online]. Available: http://dx.doi.org/10.1016/j.sigpro.2005.11.007

[15] 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, Feb. 2014. [Online]. Available: http://dx.doi.org/10.1016/j.jneumeth.2013.10.018

[16] S. Haufe, R. Tomioka, G. Nolte, K.-R. Muller, and M. Kawanabe, "Modeling Sparse Connectivity Between Underlying Brain Sources for EEG/MEG," *Biomedical Engineering, IEEE Transactions on*, vol. 57, no. 8, pp. 1954–1963, Aug. 2010. [Online]. Available: http://dx.doi.org/10.1109/tbme.2010.2046325

[17] P. A. Valdes-Sosa, "Spatio-temporal autoregressive models defined over brain manifolds." *Neuroinformatics*, vol. 2, no. 2, pp. 239–250, 2004. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/15319519

[18] J. M. Sanchez-Bornot, E. Martinez-Montes, A. Lage-Castellanos, M. Vega-Hernandez, and P. A. Valdes-Sosa, "Uncovering Sparse Brain Effective Connectivity: a Voxel-Based Approach Using Penalized Regression," *Statistica Sinica*, vol. 18, pp. 1501–1518, 2008.

[19] 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 of London. Series B, Biological sciences*, vol. 360, no. 1457, pp. 969–981, May 2005. [Online]. Available: http://dx.doi.org/10.1098/rstb.2005.1654

[20] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, Feb. 2006. [Online]. Available: http://dx.doi.org/10.1111/j.1467-9868.2005.00532.x

[21] Y. Nardi and A. Rinaldo, "On the Asymptotic Properties of The Group Lasso Estimator in Least Squares Problems," *Electron. J. Statist.*, vol. 2, pp. 605–633, 2008. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.143.177

[22] J. Huang and T. Zhang, "The Benefit of Group Sparsity," Mar. 2009. [Online]. Available: http://arxiv.org/abs/0901.2962

[23] S. Haufe, G. Nolte, K.-R. Mueller, and N. Kraemer, "Sparse Causal Discovery in Multivariate Time Series," Jan. 2009. [Online]. Available: http://arxiv.org/abs/0901.2234

[24] P. A. Valdes-Sosa, A. Roebroeck, J. Daunizeau, and K. Friston, "Effective connectivity: influence, causality and biophysical modeling." *NeuroImage*, vol. 58, no. 2, pp. 339–361, Sep. 2011. [Online]. Available: http://dx.doi.org/10.1016/j.neuroimage.2011.03.058

[25] D. P. Wipf and B. D. Rao, "An Empirical Bayesian Strategy for Solving the Simultaneous Sparse Approximation Problem," *IEEE Transactions on Signal Processing*, vol. 55, no. 7, pp. 3704–3716, Jul. 2007. [Online]. Available: http://dx.doi.org/10.1109/tsp.2007.894265

[26] Y. Qi, D. Liu, D. Dunson, and L. Carin, "Multi-task Compressive Sensing with Dirichlet Process Priors," in *Proceedings of the 25th International Conference on Machine Learning*, ser. ICML '08. New York, NY, USA: ACM, 2008, pp. 768–775. [Online]. Available: http://dx.doi.org/10.1145/1390156.1390253

[27] S. D. Babacan, S. Nakajima, and M. N. Do, "Bayesian Group-Sparse Modeling and Variational Inference," *IEEE Transactions on Signal Processing*, vol. 62, no. 11, pp. 2906–2921, Jun. 2014. [Online]. Available: http://dx.doi.org/10.1109/tsp.2014.2319775

[28] P. J. Garrigues, B. A. Olshausen, and H. W. Neuroscience, "Group sparse coding with a laplacian scale mixture prior," in *Zemel,R.,andCulotta,A.,editors,Advances in Neural Information Processing Systems*, 2010, pp. 676–684. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.230.8660

[29] A. Lee, F. Caron, A. Doucet, and C. Holmes, "A Hierarchical Bayesian Framework for Constructing Sparsity-inducing Priors," Sep. 2010. [Online]. Available: http://arxiv.org/abs/1009.1914

[30] W. D. Penny and S. J. Roberts, "Bayesian multivariate autoregressive models with structured priors," *Vision, Image and Signal Processing, IEE Proceedings* -, vol. 149, no. 1, pp. 33–41, Feb. 2002. [Online]. Available: http://dx.doi.org/10.1049/ip-vis:20020149

[31] L. Harrison, W. D. Penny, and K. Friston, "Multivariate autoregressive modeling of fMRI time series." *NeuroImage*, vol. 19, no. 4, pp. 1477–1491, Aug. 2003. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/12948704

[32] D. H. Lobato, J. M. H. Lobato, and P. Dupont, "Generalized Spike-and-slab Priors for Bayesian Group Feature Selection Using Expectation Propagation," *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 1891–1945, Jan. 2013. [Online]. Available: http://portal.acm.org/citation.cfm?id=2567724

[33] J. Hernández-Lobato, D. Hernández-Lobato, and A. Suárez, "Expectation propagation in linear regression models with spike-and-slab priors," *Mach. Learn.*, vol. 99, no. 3, pp. 437–487, Jun. 2015. [Online]. Available: http://dx.doi.org/10.1007/s10994-014-5475-7

[34] T. P. Minka, "Expectation Propagation for Approximate Bayesian Inference," in *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, ser. UAI '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 362–369. [Online]. Available: http://portal.acm.org/citation.cfm?id=647235.720257

[35] J. Riihimäki, P. Jylänki, and A. Vehtari, "Nested Expectation Propagation for Gaussian Process Classification with a Multinomial Probit Likelihood," Jul. 2012. [Online]. Available: http://arxiv.org/abs/1207.3649

[36] H. Nickisch and C. Guestrin, "Approximations for Binary Gaussian Process Classification," *Journal of Machine Learning Research*, vol. 9, pp. 2035–2078, Oct. 2008. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.165.8505

[37] S. Nishimoto, A. T. Vu, T. Naselaris, Y. Benjamini, B. Yu, and J. Gallant, "Gallant lab natural movie 4t fmri data," 2014. [Online]. Available: http://dx.doi.org/10.6080/K00Z715X

[38] S. Nishimoto, A. T. Vu, T. Naselaris, Y. Benjamini, B. Yu, and J. L. Gallant, "Reconstructing Visual Experiences from Brain Activity Evoked by Natural Movies," *Current Biology*, vol. 21, no. 19, pp. 1641–1646, Oct. 2011. [Online]. Available: http://dx.doi.org/10.1016/j.cub.2011.08.031

[39] M. Morf, A. Vieira, D. T. L. Lee, and T. Kailath, "Recursive Multichannel Maximum Entropy Spectral Estimation," *Geoscience Electronics, IEEE Transactions on*, vol. 16, no. 2, pp. 85–94, Apr. 1978. [Online]. Available: http://dx.doi.org/10.1109/tge.1978.294569

[40] H. Vinod, "A survey of ridge regression and related techniques for improvements over ordinary least squares," *The Review of Economics and Statistics*, vol. 60, no. 1, pp. 121–31, 1978. [Online]. Available: http://EconPapers.repec.org/RePEc:tpr:restat:v:60:y:1978:i:1:p:121-31

[41] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The Balanced Accuracy and Its Posterior Distribution," in *Pattern Recognition (ICPR), 2010 20th International Conference on*. IEEE, Aug. 2010, pp. 3121–3124. [Online]. Available: http://dx.doi.org/10.1109/icpr.2010.764

[42] G. Deshpande, K. Sathian, and X. Hu, "Effect of hemodynamic variability on Granger causality analysis of fMRI." *NeuroImage*, vol. 52, no. 3, pp. 884–896, Sep. 2010. [Online]. Available: http://dx.doi.org/10.1016/j.neuroimage.2009.11.060

[43] M. B. Schippers, R. Renken, and C. Keysers, "The effect of intra- and inter-subject variability of hemodynamic responses on group level Granger causality analyses." *NeuroImage*, vol. 57, no. 1, pp. 22–36, Jul. 2011. [Online]. Available: http://dx.doi.org/10.1016/j.neuroimage.2011.02.008

[44] A. K. Seth, P. Chorley, and L. C. Barnett, "Granger causality analysis of fMRI BOLD signals is invariant to hemodynamic convolution but not downsampling," *NeuroImage*, vol. 65, pp. 540–555, Jan. 2013. [Online]. Available: http://dx.doi.org/10.1016/j.neuroimage.2012.09.049

Chapter 8

Supervised Casual Graph Estimation

Supervised Causal Graph Estimation

Danilo Benozzo, Emanuele Olivetti, Paolo Avesani,

***Abstract*—Brain effective connectivity aims to detect causal interactions between distinct brain units and it is typically studied through the analysis of direct measurements of the neural activity, e.g. magneto/electroencephalography (M/EEG) signals. The literature on methods for causal inference is vast. It includes model-based methods in which a generative model of the data is assumed and model-free methods that directly infer causality from the probability distribution of the underlying stochastic process. Here, we firstly focus on the model-based methods developed from the Granger criterion of causality, which assumes the autoregressive model of the data. Secondly, we introduce a new perspective, that looks at the problem in way that is typical of the machine learning literature. Indeed, we formulate the problem of causality detection as a supervised learning task by proposing a classification-based approach. A classifier is trained to identify causal interactions in the context defined by the chosen model and according to the adopted feature space. In this paper, we are interested in comparing this classification-based approach with the standard Geweke measure of causality in the time domain through simulation study. Thus, we customized our approach to the case of a MAR model with a feature space which contains causality measures based on the idea of precedence and predictability in time. Two variations of the supervised method are proposed and compared to a standard Granger causal analysis method. As evidence of the efficacy of the proposed method, in addition to the results of the simulations, we report the details of our submission to the causality detection competition of Biomag2014, where the proposed method reached the 2nd place. Moreover, as empirical application, an example with neural recordings is provided.**

***Index Terms*—causal inference, brain effective connectivity, Granger causality, machine learning, Geweke measure in time, causal interaction classification**

I. Introduction

A main part of neuroscience research is concerned with brain connectivity and aims to investigate the pattern of interactions between distinct units within the brain [1]. The concept of brain units is strongly related to the level of the adopted scale. Thus, brain connectivity can be studied from the microscopic level of single synaptic connections to the macroscopic level of brain regions. Moreover, depending on the type of interactions of interest, brain connectivity is divided into *structural*, *functional* and *effective* connectivity. In the first case the connectivity patterns are referred to the anatomical links i.e. the neural pathways. In the second case, to the statistical dependencies between brain activity in different units. In the last case, the connectivity patterns are referred to the causal interactions between them [2]. In particular, effective connectivity provides information about the direct influence that one or more units exert over another and aims to establish causal interactions among them [3].

D. Benozzo, E. Olivetti and P. Avesani are with the NeuroInformatics Laboratory (NILab), Bruno Kessler Foundation, Trento, Italy and with the Center for Mind and Brain Sciences (CIMeC), University of Trento, Italy.

Electrophysiological signals are among the most suitable ones for studying effective connectivity. First, because they directly measure neuronal activity, even though at an aggregated level. Second, because their temporal resolution is compatible with the processing time at the neuronal level, that is in the order of milliseconds [4]. These signals can be measured with invasive or non-invasive methods. Invasive methods allow a high quality and spatially precise acquisition, by implanting electrodes on the brain. On the other side, non-invasive techniques such as magneto- and electroencephalography (M/EEG) are widely used because of the high sampling frequency and, by means of source reconstruction techniques, they provide increased signal-to-noise ratio and spatial resolution [5].

The interest in studying causal interactions from neuroimaging data is not only limited to effective connectivity but it has a more general scope. The original definition of effective connectivity provided in [3], refers to the directed influences that neuronal populations in one brain area exert on those in another one. Thus an estimator of effective connectivity should consider the physiological structure and dynamics of the system [6]. This constraint is particularly demanding since it means modeling the underlying physical processes. To overcome such issue, a relaxed version of effective connectivity was introduced in [7] under the name of *causal connectivity*. Causal connectivity refers to a causality measure that infers the causality structure without requiring it to be representative of the underlying neuronal network. The term *causality analysis* is commonly used when studying the direct interactions among brain signals. As highlighted in [8], a causality analysis may have different meanings. Its purpose could be to infer the existence of a direct causal connection, thus the estimate of the so-called causal structure or (binary) causal graph [9]. A different goal is to study the mechanism underlying a causal connection. This means focusing on how a causal connection is physiologically implemented. And a third question concerns the quantification of the interaction, thus it requires both an appropriate modeling of the dynamics and a clear understanding of what the causal effect actually means, see [10].

In this work, we focus on the problem of inferring the binary causal graph from a given set of time series. This means that our purpose is to establish the existence of causal interactions without necessarily considering the underlying mechanism and quantification issues.

A. Approaches for causal inference

In the literature, each method of causal inference is based on a specific *causality criterion* from which a measure of causality is derived [11]. A criterion of causality defines

which condition has to be satisfied in order to establish that two processes are causally interacting, or not. Given a certain criterion and according to how it is formulated, different measure of causality can be developed. There are cases in which the measure is defined by assuming a model for the underling process of data generation, the so-called parametric formulations of the criterion. Or in case of a model-free approach, the formulation is said to be non-parametric. Figure 1 summarizes the main blocks of these two approaches and introduces the main blocks of the alternative approach that is proposed in this paper, which is called the *parametric supervised approach*. The figure is horizontally divided in three parts, one for each approach. They all start by requiring a criterion of causality and a multivariate time series, as input dataset. And they all end with an estimate of the casual graph of the input dataset.

In the parametric approach, a criterion of causality is chosen and then, according to it, a model of the generative process is assumed and a measure for causality is defined. Commonly, the computation of the causality measure requires the identification of the model, which, in general, is not trivial [12]. Moreover, to obtain the causal graph from the computed measures, the significance of the non-zero values needs to be tested. This can be done, for example, by means of bootstrap techniques, or by knowing the actual distribution under the null hypothesis.

In the non-parametric approach, given a criterion of causality, its definition of causal interaction is formulated in terms of equations between probability distributions. Afterwards, a metric is adopted in the information-theoretic framework in order to test whether the equality holds [13], [14].

Differently from the parametric and non-parametric approach, here we propose a novel direction to attack the problem of detecting causality, which we call *supervised parametric* approach. The supervised approach is based on machine learning techniques and, specifically, on learning from examples. Each example comprise a multivariate time series together with their true causal structure. The idea of proposing causal inference as a learning theory problem was first presented in [15], where the authors adopted a supervised approach for bivariate causal inference with the use of kernel mean embeddings for feature mapping. Here, the same idea of a supervised detection of causal interactions is used but with a different implementation. Moreover, we specifically target the context of time series analysis. In our variant, the model is not used to derive a measure but to generate a dataset of multivariate time series together with their actual causal graphs. The purpose of this dataset is to be used as *train set* for a classification algorithm, aimed to predict the causal graph of future multivariate time series.

A consequence of the proposed approach is that we need to build a feature space in which to represent the dataset, such that the specific aspects of the chosen causality criterion are represented. Moreover, it is interesting to notice that model and feature space do not need to derive from the same causality criterion. This means that the proposed supervised approach allows to disentangle the mechanism of data generation from the criterion used to describe the causal structure.

In this work, the proposed supervised parametric approach is compared with the standard parametric formulation. For this reason, we refer to the standard parametric approach as to the *unsupervised parametric* one. In the context of the Granger criterion of causality [16], we conduct the comparison through a simulation study. Granger causality is the most adopted criterion for causal inference in brain recordings [17] and it is based on the assumptions of precedence and predictability of the cause with respect to its effect. Precedence means that a cause has to temporally precede its effect. Predictability is referred to the conditional dependence that exists between the past of the causes and the future of the effect, conditioned on the past of the effect itself.

B. Causality measures based on the Granger criterion

In the following, we provide a brief summary of the most important measures of causality that have been developed from the Granger criterion, both for the non-parametric and parametric cases.

For the non-parametric approach, a widespread causality measure is transfer entropy, which compares the probability distributions between the candidate effect and the past of the candidate cause, under the hypothesis of independence [18]–[20]. Specifically, transfer entropy computes the Kullback-Leibler divergence between the probability distribution of the candidate effect conditioned on it own past and the same effect conditioned also the past of the candidate cause. By definition, this measure is non-negative and zero only when the two distributions are equal. Moreover, the fact that KL-divergence does not consider any specific statistical moment of a given order, is particularly suited for detecting non-linear interactions. Beyond transfer entropy, other non-parametric measures have been proposed [21], such as the measure based on Fisher information.

The parametric representation of the Granger criterion assumes a linear autoregressive model of the process. This assumption refers to how time series are interacting with each other, without explicitly modeling the physical mechanism of generation. The autoregressive representation has led to different formulations of measures of causal interaction. The temporal formulation tests the presence of causality by comparing the residual variances of the effect in which the candidate cause is initially excluded vs. when it is included, during model identification. The causal measure is defined as the natural logarithm of the ratio of the residual variances, that we refer to as the *Geweke measure in time domain*. A meaningful reduction of the residual variance when the candidate cause is included in model identification means a better model for the effect. In such case, the time series evaluated as possible cause is said to *Granger cause* the time series evaluated as effect [22]. It has been proven that this measure of causality is a test of Granger causality on the first moment statistic of the underlying probability distributions [23], since it is based on the linear assumption of the process. This is in contrast with transfer entropy where, by definition, the whole probability distribution of the processes is considered [24].

The autoregressive parametric formulation of the Granger criterion was also implemented in the spectral domain. It was

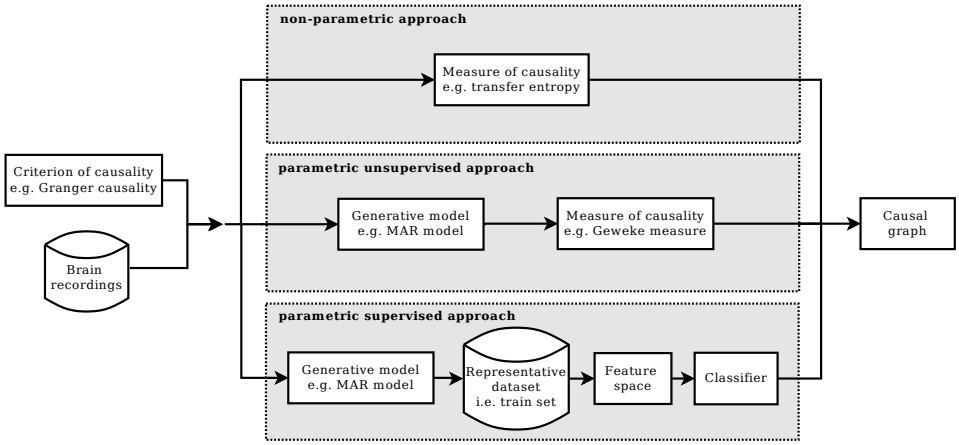


Fig. 1: Given a criterion of causality, the estimation of causality structure can be implemented in three different ways: the standard non-parametric approach (top), the parametric one (mid) and the proposed parametric supervised one (bottom).

introduced in [25] and named *Geweke spectral measure* of Granger causality. In the bivariate case, the Geweke spectral measure from x and y at the frequency ω, is defined as the natural logarithm of the ratio of the power spectrum of y computed considering the possible contribution of x and the power spectrum of y computed alone, in both cases evaluated at ω. It is interpreted as the portion of the power spectrum associated with the residuals that do not take into account the presence of y [26]. The Geweke spectral measure does not have its equivalent formulation in the information-theoretic framework. As shown in [26], the lack of a temporal separation between the past and the future of the involved processes is what defines a spectral formulation of a parametric formulation. Differently, in the non-parametric formulation, a spectral measure is not available, because no way to avoid temporal separation has been proposed yet.

Other examples of causal measures developed in the spectral domain are the Partial Directed Coherence (PDC) [27] and the Direct Transfer Function (DTF) [28]. Both were initially developed under the assumption of identity matrix as covariance matrix of the innovation process and then generalized in [29], where they are named the information PDC (iPDC) and the information DTF (iDTF). Both are defined as a coherence measure between two processes thus they have an interpretation in term of mutual information rate. Moreover, both are measures to test for Granger causality, but only in the case of DTF, a direct connection between the bivariate Geweke spectral measure and the bivariate iDTF exists. iPDC assumes an autoregressive model for the process while iDTF starts with the moving average representation of the autoregressive model.

In the neuroscience domain, the multivariate extension of the causality measures introduced so far has great importance [30]. In the case of the bivariate iPDC and iDTF, the multivariate extension are straightforward [31]. Also the Geweke measure in time domain has a direct multivariate extension from the bivariate case, by conditioning on the processes that are not included in the pair [32]. Less immediate is the extension of the spectral representation: for a detailed explanation see [33].

C. Proposal

The aim of this work is to investigate the proposed supervised formulation by adopting a parametric model of the Granger criterion of causality. We propose a simulation study in the context of the autoregressive model, specifically in the time domain. With these ingredients, it is possible have a fair comparison against the standard conditional Geweke measure in time domain. Across the experiments, we compare the proposed method against a standard Granger causal analysis (GCA) method [34].

The proposed approach is analyzed in a series of experiments that are grouped in two parts. What differs between them is the generative process used for the training and for the test/prediction phase. In the first group, the model is the same for the train and the test phase. The first group is meant to evaluate the proposed approach under the three main aspects of the method: the generative model, the feature space and the classification task. In the second group, the generative model differs between train set and test set. This case is quite common in practical cases, because the recorded signals may not fully respect the assumptions of the generative model assumed for the analysis.

In addition, we report the details of the solution computed with the supervised method that we submitted to the Biomag2014 Causality Challenge (Causal2014)b[1], which reached the second place of the ranking. Such competition adopted an autoregressive model as generative process to simulate brain signals. The model generated a 3-dimensional multivariate time series, given a randomly generated causal

[1]http://www.biomag2014.org/competition.shtml , see "**Challenge 2**: Causality Challenge".

graph[2]. The competition distributed a large set of these multivariate time series and the task was to reconstruct their causal graphs.

In the second part of the experiments, we introduced a mismatch between the generative process of the training phase and the process of the prediction phase. The purpose of studying such situation is to assess how strong is the bias of the generative model, i.e. the one used to create the train set, when predicting data coming from a (partly) different process. Two different cases are analyzed in the second part: one with simulated datasets and the second with neural recordings from rats.

II. Materials

In this section, we describe the multivariate autoregressive model (MAR) used in our simulations and then the neural recordings used for testing the proposed method in a real setting.

A. The MAR model

The final output of the MAR model is the multivariate time series $\mathbf{X} = \{X(t), t = 0, 1, \ldots, N-1\}, X(t) \in \mathbb{R}^{M\times 1}$ that is defined as the linear combination of two M-dimensional multivariate time series $\mathbf{X_s}$ and $\mathbf{X_n}$

$$\mathbf{X} = (1-\gamma)\mathbf{X_s} + \gamma\mathbf{X_n} \tag{1}$$

$\mathbf{X_s}$ carries the causal information, $\mathbf{X_n}$ represents the noise corruption and $\gamma \in [0, 1]$ tunes the signal-to-noise ratio. The choice of this formulation of the MAR model, with additive noise included, is motivated by the facts that Granger metrics are strongly affected by both uncorrelated and linearly mixed additive noise [35] and because it was also adopted in the Causal2014 competition. Each time point of $\mathbf{X_s}$ and $\mathbf{X_n}$ is computed by following the MAR model

$$\begin{aligned} X_s(t) &= \sum_{\tau=1}^{p} A_s^{(\tau)\top} X_s(t-\tau) + \varepsilon_s(t) \\ X_n(t) &= \sum_{\tau=1}^{p} A_n^{(\tau)\top} X_n(t-\tau) + \varepsilon_n(t) \end{aligned} \tag{2}$$

where p is the order of the MAR model and represents the maximal time lag, $\varepsilon_s(t)$ and $\varepsilon_n(t)$ are realizations from a M-dimensional standard normal distribution and $A_s^{(\tau)}, A_n^{(\tau)} \in \mathbb{R}^{M\times M}, \tau = 1, \ldots, p$, are the coefficient matrices modeling the influence of the signal values at time $t-\tau$ on the current signal values, i.e. at time t. The coefficient matrices $A_s^{(\tau)}$ are involved in the process of causal-informative data generation. They are computed by multiplying the non-zero elements of the $M \times M$ binary matrix A with uniformly distributed random numbers. In essence, A is called causal configuration matrix and represents the causal graph that leads the MAR model. Specifically $A_{ij} = 1$ means signal i causes the signal j. On the other hand, coefficient matrices $A_n^{(\tau)}$ lead the noisy part of the signals and they are obtained by randomly generating p diagonal matrices. After that, if both sets of matrices $A_s^{(\tau)}$ and $A_n^{(\tau)}$ fulfill the stationarity condition, each time point of $\mathbf{X_s}$ and $\mathbf{X_n}$ is generated by Equation 2.

[2]Represented as a 3×3 binary matrix.

B. Neural recording dataset

The neural recording data that have been used for the real application experiment, belong to the *hc-3 dataset* [36], [37]. The dataset and related details on the acquisition are available online at https://crcns.org/data-sets/hc/hc-3. Neural time series were recorded from rats while they were performing multiple behavioral tasks. We only used local field potentials from session eco013.156 of three specific shank probes, i.e. the ones associated to the Cornu Ammonis (CA1) and the entorhinal cortex (EC3 and EC5). Each shank has 8 recoding sites. Signals were low pass filtered at 140 Hz, down-sampled at 600 Hz and epoched into non-overlapping segments of 5s duration. Moreover, we averaged across recording sites in each shank. Our final dataset contains 102 trials each of 3 time series with 5s length associated to the three brain areas (CA1, EC3 and EC5). In order to quantify the accuracy of the evaluated methods, the true causal configuration matrix was defined by assuming the following chain of interactions: EC3→CA1→EC5, as in [38].

III. Methods

In this paper, we propose a parametric supervised approach to the problem of causal inference. The idea is to define the causal inference in a supervised machine learning framework, in which a classifier learns how to discriminate among a set of defined classes, i.e. causal configurations, though a training phase. The approach is *parametric* because a model of the generative process is assumed and used to generate examples for the training phase. In details, there are two main ingredients to handle the problem in a parametric supervised way: the first is the definition of a model of the stochastic process underling the time series and the second is the definition of a feature space able to capture the causal relationships of a given set of time series. The choice of the model is a step in common with all other parametric criteria for causal inference. The difference is that, in our case, the model is used for the generation of the train set instead of the formulation of a measure of causality. In order to compare the supervised framework with the Geweke measure of causality in time domain, we instantiated our method with the MAR model. Moreover, we designed a feature space based on the idea of predictability and precedence in time, as in the Geweke measure [3]. In the following we report all the details of this procedure.

A. Data generation and causal configuration

The train dataset, that is class-*labeled* and denoted as L, is generated considering the total number of causal graphs that can be produced by a given number of time series. In a general setting, each trial $\mathbf{X}$ is composed by M time series and the final goal of causal inference is to estimate its $M \times M$

[3]Python implementation at: https://github.com/danilobenozzo/supervised_causality_detection.git

binary configuration matrix A. Thus, there are $M(M-1)$ free binary parameters and $2^{M(M-1)}$ possible causal configuration matrices [4]. Considering that $\mathbf{L}$ must be representative of the entire population of configurations, it will be generated so that multiple trials are included for each possible causal graph.

B. Classification schema: MBC and CBC

Here we describe two versions of the parametric supervised method. In the first, the entire causal configuration matrix A is considered the class label of the trial. This choice means that one classifier has to be trained to discriminate among $2^{M(M-1)}$ classes. We will refer at this solution as the *matrix-based* classification (MBC) method. In the second version of the parametric supervised method, each cell of the configuration matrix is analyzed independently from other cells. Since each cell can be only 0 or 1, then the whole problem of predicting the causal configuration is transformed into $M(M-1)$ binary classifications problems, one for each cell. We call this approach the *cell-based* classification (CBC).

C. Definition of the feature space

The feature space is defined on the same assumptions done in the case of the autoregressive formulation of Granger causality. Thus, each trial is mapped into a vector of measures, each based on the ability to predict the value of one time series at a given time point, i.e. the *effect*, from the past values of each possible subset of the M time series in the trial, i.e. the possible *causes*. We call the pair, made by causes and effect, *causality scenario*. In other words, chosen one of the M time series as the effect in the causality scenario, the related possible causes are all the subsets that can be formed from the whole set of time series. For M time series, the number of scenarios is $\sum_{i=1}^{M} \binom{M}{i} M = (2^M - 1)M$, by using the binomial theorem. In Table I, we report the causality scenarios in case $M = 3$. Thus, the possible causality scenarios are 7 for each $x_i(t), i = 0, 1, 2$, i.e. time series that defines a trial, so 21 causality scenarios in total.

	Causes	Effect
1	$x_0(t)$	$x_i(t)$
2	$x_1(t)$	$x_i(t)$
3	$x_2(t)$	$x_i(t)$
4	$x_0(t), x_1(t)$	$x_i(t)$
5	$x_0(t), x_2(t)$	$x_i(t)$
6	$x_1(t), x_2(t)$	$x_i(t)$
7	$x_0(t), x_1(t), x_2(t)$	$x_i(t)$

TABLE I: For each effect $x_i(t)$ and $M = 3$, we report the 7 possible causality scenarios.

For each causality scenario, a plain linear regression problem is built by selecting, as dependent variable, the time points from the signal in the *effect* column. Each of these dependent variables has a regressor vector composed by the p previous time points selected from the signals in the *causes* column, where p is the order of the MAR model, see Section II-A. Table II shows how the regression problems are defined when

[4]The diagonal is not relevant since by definition time series are autoregressive.

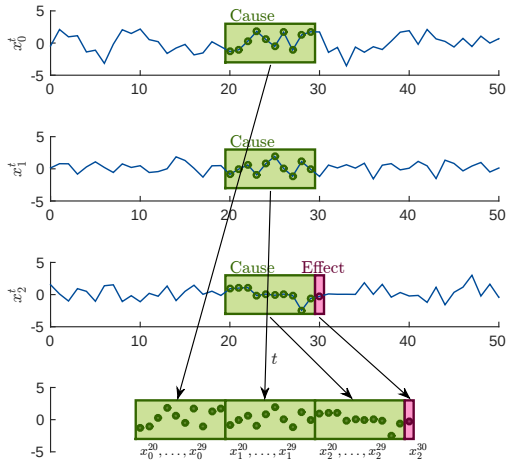


Fig. 2: Example of how the sample associated at the time point $t = 30$ is built in order to form the input of the last regression problem in Table II, for the case $i = 2$ and $p = 10$.

$M = 3$, by specifying from which time series and time points, regressors and dependent variables are extracted. In the following, in order to simplify the notation, we will use x_i^t instead of $x_i(t)$, $i = 0, 1, 2$ and $t \in \mathbf{T}, \mathbf{T} \subseteq \{p, \ldots, N-1\}$. Figure 2 explains how, for the specific time point $t = 30$ and for $p = 10$, the input of the regression problem is built for the last causality 7 of Table II and $i = 2$: $\{x_0, x_1, x_2\} \rightarrow x_2$. Finally, the regression problem of each causality scenario is scored, by common metrics like the means squared error. Such scores are used as feature in the feature space representation of the training set L.

D. Relationship with the Geweke measure

As summarized in Table I, the feature space is defined by exploiting all the possible causality scenarios among a set of M time series. Differently, in the *bivariate* case, the Geweke measure separately tests for each pair (x_i, x_j) the cases of $x_i \rightarrow x_j$ and $x_j \rightarrow x_i$. In terms of the scenarios described above, the bivariate evaluation of $x_i \rightarrow x_j$ corresponds to the cases $x_j \rightarrow x_j$ and $\{x_i, x_j\} \rightarrow x_j$. This means that, when considering 3 or more time series, the Geweke measure would consider only a pairwise analysis.

Similarly, the *conditional*-bivariate implementation of the Geweke measure tests the causal interaction by including in the set of causes of each causality scenario the $M - 2$ time series that are not in the pair under analysis.

In the analysis of the proposed method, we will also consider the subsets of feature space that corresponds to the bivariate and conditional-bivariate cases, by removing the scenarios not included in those cases. For clarity, we call the two reduced features spaces as *pairwise* (pw) and *conditional-pairwise* (c-pw). In both cases, given M time series and selected one as effect, its possible causes define $M - 1$ causality scenarios plus the causality scenario that involves

Regressors (causes)	Dependent variable (effect)
$[x_0^{t-p}, \ldots, x_0^{t-1}]$	x_i^t
$[x_1^{t-p}, \ldots, x_1^{t-1}]$	x_i^t
$[x_2^{t-p}, \ldots, x_2^{t-1}]$	x_i^t
$[x_0^{t-p}, \ldots, x_0^{t-1}, x_1^{t-p}, \ldots, x_1^{t-1}]$	x_i^t
$[x_0^{t-p}, \ldots, x_0^{t-1}, x_2^{t-p}, \ldots, x_2^{t-1}]$	x_i^t
$[x_1^{t-p}, \ldots, x_1^{t-1}, x_2^{t-p}, \ldots, x_2^{t-1}]$	x_i^t
$[x_0^{t-p}, \ldots, x_0^{t-1}, x_1^{t-p}, \ldots, x_1^{t-1}, x_2^{t-p}, \ldots, x_2^{t-1}]$	x_i^t

TABLE II: Description of how the 21 linear regression problems are defined for each trial. x_i^t, $i = 0, 1, 2$ and $t \in \mathbf{T}, \mathbf{T} \subseteq \{p, \ldots, N-1\}$, are the three time series of a trial.

only the effect itself, i.e. $x_j \rightarrow x_j$. Thus, we evaluate M causality scenarios for each effect, and M^2 in total. In our example of $M = 3$, the number of dimensions is 9, instead of the 21.

E. Evaluation metrics

In the experiments described in Section IV, the ability to identify the correct causal configuration on simulated and real datasets, will be quantified in terms of receiver operating characteristic (ROC) curve and the related area under the curve (AUC). In this way, the obtained results will not be biased by the possible different cost of false discovery that may change in different applications.

The computation of the ROC curve in the cases of the standard Granger causality analysis (GCA, see [34]) and cell-based classification (CBC) is straightforward, because GCA is a conditioned pair-wise method and CBC predicts the single cells of the causality matrix. The ROC curve can then be computed from the false positive (FP) rate and the true positive (TP) rate obtained by varying the classification threshold[5] and by averaging over all cells and all trials.

In the case of matrix-based classification (MBC), the classification problem is multiclass and the ROC curve cannot be obtained in a straightforward way, in general. Nevertheless, in our specific case, each predicted causal matrix is a binary matrix, as in the case of CBC. The only difference is that, with MBC, all entries of the matrix are jointly predicted instead of being individually predicted each by a different classifier, as in CBC. Anyway, by jointly varying the classification threshold in all entries of the matrix, we can compute the ROC curve for MBC, allowing a fair comparison with CBC and GCA.

IV. Experiments

The purpose of our empirical analysis is to compare the proposed supervised methods, described in Section III, against the best practice in the literature, which is based on an unsupervised estimate of the parameters of the MAR model. The comparison is performed mainly with synthetic data where the ground truth of effective connectivity is known in advance, by design. Additionally, on real data, we investigate the behavior of the supervised approach when the underlying exact generative model is not known in advance. To conclude, we also report the empirical investigation proposed by the Causal2014 challenge[6].

A. Data generation process and feature space

Before describing each experiment, we provide details on the initialization of the MAR model to generate the dataset $\mathbf{L}$ and on how to create and improve the feature space described in Section III-C. The parameters of the MAR model were set as $p = 10$, $N = 6000$ and $M = 3$. Regarding the parameter γ, since the presence of additive noise affects the performance of a Granger-based metric, we generated two versions of the $\mathbf{L}$ dataset. One version that we call $\mathbf{L}_{\text{MAR}}$, contains only the autoregressive component and no noise corruption. This practically means keeping $\gamma = 0$ in Equation 1. In a second version, with explicit noise corruption, γ is picked uniformly at random for each trial. We refer to this last dataset as $\mathbf{L}$. Given this setting, there are $2^6 = 64$ possible causal graphs / configurations. 1000 trials were generated for each configuration, thus in total 64000 trials comprised $\mathbf{L}_{\text{MAR}}$ and $\mathbf{L}$.

As explained in Section III-C, as feature space we computed two regression metrics: the mean square error and the coefficient of determination r^2. Both were included because we noticed a significant improvement in the cross-validated score, although, intuitively, they could seem redundant. Additionally, we included an estimate of the Granger causality coefficients [7]. As a further step, we enriched the feature vector by applying standard feature engineering techniques, like simple basis functions. These consisted in extracting the 2nd power, 3rd power and square root of the previously defined features, together with the pairwise product of all features. Adding extracted features was motivated by the need to overcome the limitation of the adopted linear classifier, see [39].

B. Experiments with the same process of data generation

The experiments with simulated data were designed according to the three main components of the supervised approach: (i) the generative model, (ii) the encoding of the signals into the feature space and (iii) the use of a classification task.

The first experiment aimed to investigate the effect of including additive noise to the data generation process. Both the unsupervised (GCA) and supervised methods were applied

[5] We assume to use classifiers that produce a classification score, like the probability of having a causal interaction.

[6] http://www.biomag2014.org/competition.shtml , see "**Challenge 2**: Causality Challenge".

[7] See `GrangerAnlayzer` in NiTime, http://nipy.org/nitime

to the two datasets $\mathbf{L}_{\text{MAR}}$ and $\mathbf{L}$. For the implementation of GCA, we adopted the toolbox proposed in [34]. For the supervised approach, after the mapping of the datasets to the proposed feature space, the logistic regression classifier [8], with ℓ_2 regularisation, was applied in a 5-folds cross-validation framework.

The second experiment aimed to characterize the properties of the feature space proposed in Section III-C, that we call *complete* feature space, and to compare it with its pairwise (pw) and conditional-pairwise (c-pw) versions described in Section III-D. Such restricted/reduced feature spaces were introduced to mimic the Geweke measure, which addresses the bivariate case. The aim is to understand the gain of introducing the *complete* feature space that accounts also for the multivariate case.

The third experiment considered the two alternative schema to formulate the classification task: the matrix-based classification (MBC), which jointly predicts all entries of the causal matrix, and the cell-based classification (CBC), for which each matrix cell refers to a different binary classifier, see Section III-B. Since $M = 3$, in the case of MBC we trained one classifier to predict among 64 different classes, one for each possible causal configuration matrix. In the case of CBC, 6 binary classifiers were trained, one for each cell of the causal matrix. Both versions were applied to the two simulated datasets $\mathbf{L}_{\text{MAR}}$ and $\mathbf{L}$.

As an additional evaluation of the supervised approach, here we report the detail of our submission to the Causality2014 challenge. In this setting we know in advance the generative model, i.e. the MAR model, but the ground truth of the causal graph of each trial is unknown. We used the $\mathbf{L}$ dataset as train set with the MBC method with the complete feature space. The posterior probabilities computed by the logistic regression classifier were converted into predicted classes considering the costs provided by the competition for true positives (+1) and false negatives (-3).

C. Mismatch between generative processes

In this experiment, we artificially introduced a mismatch between the generative model of the train set and the actual process of signal generation. This is a frequent scenario in practical cases, because generative models are only approximations of the real physical process creating the data. For this reason, we wanted to compare the proposed supervised approach with respect to the standard analysis under such scenario. In practice, we applied CBC to the $\mathbf{L}$ dataset after training it on the $\mathbf{L}_{\text{MAR}}$ dataset and, as feature space, we adopted its complete version.

As a second experiment on the mismatch of the generative processes, we trained the CBC method on the $\mathbf{L}$ dataset and tested on the real neural recording dataset described in Section II-B. The experiment was repeated with different configurations, i.e. by changing the sampling frequency of the neural signals and the related model order p. As sampling frequency, we set it to 600, 800 and 1000 Hz and the model order was computed in order to have time windows of 5, 10, 15, 20 and 25 ms. For each pair of sampling frequency and model order the AUC was computed using as true causal configuration matrix the causal chain reported in Section II-B, i.e. EC3→CA1→EC5, as in [38].

[8]http://scikit-learn.org

V. Results

In this section, we report the results of the multiple experiments described in Section IV. There, we presented two groups of experiments that we report here too.

In the first group of experiments the model of data generation is exactly the same of the dataset to be analyzed. In other words, the simulated train set and the test of the supervised approach are generated with the same data generation process. The results of the first experiment, i.e. comparing GCA and the propose supervised methods on data with and without additive noise, are presented in Table III as ROC AUC scores (higher is better). As expected, with no additive noise, see row $\mathbf{L}_{\text{MAR}}$, all methods predict identically, because classification is perfectly accurate in all cases. When adding noise, i.e. row $\mathbf{L}$, the AUC score changes from 0.72 for GCA to 0.90-0.92 for the supervised methods.

The second experiment of the first group compares the different features spaces for the supervised approach. In Table III, the AUC of the complete feature space (columns CBC, MBC), of the pairwise one (column CBC pw) and of the conditional-pairwise one (column CBC c-pw), are reported. The corresponding ROC curves are illustrated in Figure 4.

The third experiment of the first group, compares our two different approaches to classification, i.e. the cell-based (CBC) and the matrix-based (MBC) ones. In Table III, columns 2 and 5, the AUC scores are reported together with those of GCA. The full ROC curve is presented in Figure 3.

The first group of experiments is concluded by the results of the Causal2014 challenge, reported in Table IV. The results are 5-fold cross-validated on the train set, because the causality matrices of the test set of the competition were not disclosed. The table reports the confusion matrices of GCA, CBC and MBC estimated on $\mathbf{L}$, following the competition guidelines.

The second group of experiments, investigates the effect of a generative model that differs from the actual generation process of the data to analyze, i.e. there is a mismatch between the two models. The first experiment, where CBC was trained on $\mathbf{L}_{\text{MAR}}$ and tested on $\mathbf{L}$, resulted in a ROC AUC score of 0.85. Despite the difference between train set and test set, the result is superior to the 0.72 obtained by GCA, see Table III. The results of the second experiments, of CBC on the neural recording dataset (see Section II-B), are reported in Table V, in terms of AUC score for different choices of the sampling frequency and order of the MAR model (p), i.e. the window width. We computed the AUC score also for GCA, obtaining chance-level results, i.e. AUC ≈ 0.5, in all cases. We observed that GCA estimated the existence of causal links in almost all cases/interactions, clearly generating a very large number of false positives. At the same time, we observed that the neural recording data have high autocorrelation and cross-correlation, which may explain such behavior.

		Pred. (GCA)	
		1	0
True	1	99.6%	0.4%
	0	80.1%	19.9%

		Pred. (CBC)	
		1	0
True	1	59.4%	40.6%
	0	2.8%	97.2%

		Pred. (MBC)	
		1	0
True	1	57.8%	42.2%
	0	2.2%	97.8%

TABLE IV: Confusion matrices of GCA, CBC and MBC on the Causal2014 dataset, taking into account for the bias for reducing the false-positives. The values are conditional probabilities given the true class, i.e. each row sums up to 1.

	GCA	CBC	CBC c-pw	CBC pw	MBC
$\mathbf{L}_{\text{MAR}}$, i.e. $\gamma = 0$	1	1	1	1	1
$\mathbf{L}$, i.e. $0 \leq \gamma \leq 1$	0.72	0.92	0.91	0.90	0.91

TABLE III: AUC values of GCA, CBC (with the complete and reduced feature spaces) and MBC on the two datasets $\mathbf{L}_{\text{MAR}}$ and $\mathbf{L}$. The standard deviation is lower than 0.0015 in all cases.

	5ms	10ms	15ms	20ms	25ms
600Hz	0.80	0.82	0.82	0.83	0.82
800Hz	0.82	0.82	0.82	0.73	0.62
1kHz	0.82	0.82	0.75	0.61	0.64

TABLE V: AUC computed by applying CBC to the empirical dataset with different sampling frequencies and time window widths.

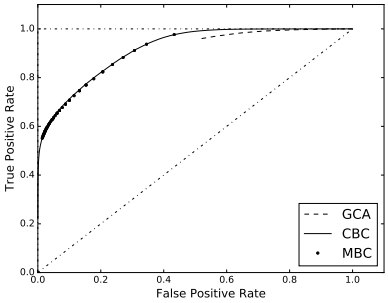


Fig. 3: ROC curves estimated on the results of the three analysed causal inference methods: Granger Causality Analysis (GCA), Cell-based Classification (CBC) and Matrix-based Classification (MBC).

VI. Discussion and conclusion

In this paper, we propose a new approach for causal inference in the framework of machine learning. Specifically, we developed a classification-based method by assuming a model for the stochastic process and a causality measure, and created a suitable feature space. Our idea is to use the model to generate a simulated dataset, representative of the problem of which we want to infer the causal interactions. Then we map this dataset into a suitable feature space. After that, a classifier is trained on the dataset in order to predict the causal graph of a future set of time series, i.e. to predict a set of binary variables. As a consequence, the causal inference is directly dependent both on the chosen generative model and on the designed feature space.

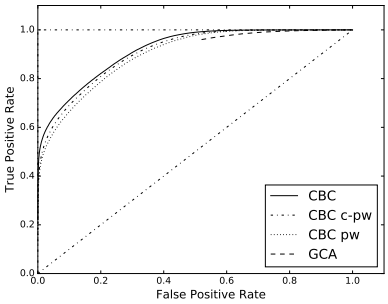


Fig. 4: ROC curves estimated on the results of CBC when applied on three different feature spaces: the complete one in contrast with the pw and c-pw ones. The ROC curve of GCA is shown as benchmark.

We put this general framework in practice, by customizing it in the case of the Geweke causal inference in time, see Section III, and, as a consequence, of the autoregressive model as the generative process of the multivariate time series. Moreover, another consequence is the assumption of precedence and predictability in time, for the identification of a causal interaction. In Section III and Section IV, we designed a feature space coherent with these assumptions.

In the experiments of Section IV, we compared the performance of different methods for causal inference, when applied to a multivariate autoregressive dataset, with and without additive uncorrelated noise. The results are shown in terms of AUC value and ROC curve, see Figure 3. The estimated AUC of each method on each dataset is reported in Table III. In absence of correlated noise, i.e. with dataset L_{MAR}, all methods perfectly predicted the correct causal configurations, which is a positive sanity check of the supervised approach. With the presence of additive noise, predicting the correct causal configuration become more difficult. In particular, we observed that GCA is more sensitive to additive noise than the supervised approaches, scoring AUC $= 0.72$, with respect to $0.90 - 0.92$ of the supervised methods. Figure 3 confirms that both the supervised methods CBC and MBC perform better than GCA. It is interesting to note that the ROC curve of GCA does not exist for false positive rate lower than 0.55. This occurs because the poor granularity of the scores of GCA does not allow to put thresholds that result in a false positive rate lower than 0.55. Specifically, GCA assigns probability 1.0 to a large amount of causal interactions that are not existent. In these result and other experiments, we observed that GCA tends to overestimate the presence of causal interactions. Differently, both CBC and MBC have much more granularity and higher AUC scores, i.e. 0.92 and 0.91 respectively. Given that both CBC and MBC operate on

the same feature space, we can conclude that a joint prediction of all causality interactions, which is what MBC provides, does not result in an advantage over the individual predictions of each interactions, which is what CBC provides.

The proposed supervised approach allows to study multivariate causal interactions. This is different from the Geweke measure, that is a conditioned pairwise method. In the supervised case all the multivariate dependencies among time series are taken into account through the causality scenarios included in the designed feature space, see Section III-C. For this reason, the proposed approach goes beyond what the pairs of cause/effect time series can give. In Figure 4 and Table III are reported the results of the analysis on the role of the proposed features space. When considering only the pairwise (CBC pw) and conditional pairwise (CBC c-pw) portions of the feature space, the AUC score is lower than the full feature space (CBC), even tough by a margin.

Considering the specific case of the Causal2014 challenge, we reported in Table IV the confusion matrices computed with GCA, CBC and MBC on the train set through cross-validation, considering the cost model defined in the competition, see Section IV-B. From this example, we clearly see that GCA provided a very large fraction of false positive, i.e. 80.1%. Differently, both CBC and MBC correctly followed the bias of the competition of reducing the number of false positives, which was 2.8% and 2.2% respectively. Our submission to the competition, with MBC [9], reached the 2nd place in the ranking, which is positive evidence that, in the case of the Geweke measure, the supervised approach is a meaningful alternative to the current state of the art unsupervised causal inference methods.

In practical cases, generative models may not accurately describe the observed data coming from neuroimaging experiments. For this reason, we wanted to test the effect of introducing a systematic change between the train set and the test set. Such change may have particularly negative impact for classification-based methods. Then, CBC was trained on L_{MAR} and then tested on L. As reported in Section V, in this case AUC dropped to 0.85, from 0.92 of the case where L was both the train set and the test set. Such result is still superior to AUC $= 0.72$, obtained with GCA. Such evidence supports the hypothesis that CBC is also robust to some violations in the assumption of the generative model.

On the neural recordings dataset introduced in Section II-B, the assumption of the MAR model may be incorrect. In Section IV, we reported that on such data GCA performed poorly, around chance-level, in all cases. This may be explained by both incorrect assumptions and by the high autocorrelation and cross-correlation in the time series. Differently from GCA, in Table V we show that CBC reaches high AUC scores, i.e around 0.82, for all sampling frequencies. We notice that, for larger time windows and higher frequencies, the AUC drops to 0.61, probably due to the increase in high frequency noise in the data. Nevertheless, it has to be noted that these results assume the validity of the causal chain EC3→CA1→EC5 that was introduced in [38].

[9]At the time of the competition we had developed only MBC and not CBC.

A. Computational Limitations

In the experiments proposed in this work, we limited the number of time series to $M = 3$. Following the explanations in Section III-B and Section III-C, this results in 64 classes, in case of MBC, or 6 binary problems, in case of CBC, and a feature space of 21 dimensions[10]. The first and the last number increase exponentially with M and the second quadratically with M. For $M = 4$, the three numbers become 4096, 12 and 60, respectively. For this reason, MBC becomes unfit to be used when $M > 3$, because the train set necessary to fit the parameters for a very large number of classes would be unfeasible to obtain and to manage. Nevertheless, even the use of CBC cannot address a large number of time series, because the feature space grows exponentially with M.

Nevertheless, it is interesting to note that the feature space proposed in this work is not bound to the generative model considered here, i.e. the MAR model. The causality scenarios defined in Section III-C are based on the causality measure, i.e. the Geweke measure. This opens interesting avenues for further research, which investigates how the inference based on the same feature space would change when different models of the generative process are used.

[10]without considering the additional feature engineering step described in Section IV-A.

REFERENCES

[1] B. Horwitz, "The elusive concept of brain connectivity." *NeuroImage*, vol. 19, no. 2 Pt 1, pp. 466–470, Jun. 2003. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/12814595

[2] V. Sakkalis, "Review of advanced techniques for the estimation of brain connectivity measured with EEG/MEG," *Computers in Biology and Medicine*, vol. 41, no. 12, pp. 1110–1117, Dec. 2011. [Online]. Available: http://dx.doi.org/10.1016/j.compbiomed.2011.06.020

[3] K. J. Friston, "Functional and effective connectivity: a review." *Brain Connectivity*, vol. 1, no. 1, pp. 13–36, 2011. [Online]. Available: http://dx.doi.org/10.1089/brain.2011.0008

[4] J.-M. M. Schoffelen and J. Gross, "Source connectivity analysis with MEG and EEG." *Human brain mapping*, vol. 30, no. 6, pp. 1857–1865, Jun. 2009. [Online]. Available: http://dx.doi.org/10.1002/hbm.20745

[5] M. J. Brookes, M. W. Woolrich, and G. R. Barnes, "Measuring functional connectivity in MEG: a multivariate approach insensitive to linear source leakage." *NeuroImage*, vol. 63, no. 2, pp. 910–920, Nov. 2012. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/22484306

[6] K. J. Friston, A. M. Bastos, A. Oswal, B. van Wijk, C. Richter, and V. Litvak, "Granger causality revisited." *NeuroImage*, vol. 101, pp. 796–808, Nov. 2014. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/25003817

[7] S. L. Bressler and A. K. Seth, "Wiener-Granger causality: a well established methodology." *NeuroImage*, vol. 58, no. 2, pp. 323–329, Sep. 2011. [Online]. Available: http://dx.doi.org/10.1016/j.neuroimage.2010.02.059

[8] D. Chicharro and A. Ledberg, "When Two Become One: The Limits of Causality Analysis of Brain Dynamics," *PLoS ONE*, vol. 7, no. 3, pp. e32 466+, Mar. 2012. [Online]. Available: http://dx.doi.org/10.1371/journal.pone.0032466

[9] M. Eichler, "A graphical approach for evaluating effective connectivity in neural systems." *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, vol. 360, no. 1457, pp. 953–967, May 2005. [Online]. Available: http://dx.doi.org/10.1098/rstb.2005.1641

[10] B. Schelter, J. Timmer, and M. Eichler, "Assessing the strength of directed influences among neural signals using renormalized partial directed coherence." *Journal of neuroscience methods*, vol. 179, no. 1, pp. 121–130, Apr. 2009. [Online]. Available: http://dx.doi.org/10.1016/j.jneumeth.2009.01.006

[11] D. Chicharro, "Parametric and Non-parametric Criteria for Causal Inference from Time-Series," in *Directed Information Measures in Neuroscience*, ser. Understanding Complex Systems, M. Wibral, R. Vicente, and J. T. Lizier, Eds. Springer Berlin Heidelberg, 2014, pp. 195–219. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-54474-3_8

[12] P. A. Valdes-Sosa, A. Roebroeck, J. Daunizeau, and K. Friston, "Effective connectivity: influence, causality and biophysical modeling." *NeuroImage*, vol. 58, no. 2, pp. 339–361, Sep. 2011. [Online]. Available: http://dx.doi.org/10.1016/j.neuroimage.2011.03.058

[13] R. Vicente, M. Wibral, M. Lindner, and G. Pipa, "Transfer Entropy–a Model-free Measure of Effective Connectivity for the Neurosciences," *J. Comput. Neurosci.*, vol. 30, no. 1, pp. 45–67, Feb. 2011. [Online]. Available: http://dx.doi.org/10.1007/s10827-010-0262-3

[14] V. Solo, "On causality and mutual information," in *Decision and Control, 2008. CDC 2008. 47th IEEE Conference on*. IEEE, Dec. 2008, pp. 4939–4944. [Online]. Available: http://dx.doi.org/10.1109/cdc.2008.4738640

[15] D. Lopez-Paz, K. Muandet, B. Schölkopf, and I. Tolstikhin, "Towards a Learning Theory of Cause-Effect Inference," May 2015. [Online]. Available: http://arxiv.org/abs/1502.02398

[16] C. W. J. Granger, "Investigating Causal Relations by Econometric Models and Cross-spectral Methods," *Econometrica*, vol. 37, no. 3, pp. 424–438, Aug. 1969. [Online]. Available: http://dx.doi.org/10.2307/1912791

[17] A. K. Seth, A. B. Barrett, and L. Barnett, "Granger Causality Analysis in Neuroscience and Neuroimaging," *The Journal of Neuroscience*, vol. 35, no. 8, pp. 3293–3297, Feb. 2015. [Online]. Available: http://dx.doi.org/10.1523/jneurosci.4399-14.2015

[18] T. Schreiber, "Measuring information transfer," *Physical review letters*, vol. 85, no. 2, pp. 461–464, Jul. 2000. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/10991308

[19] P.-O. Amblard and O. J. J. Michel, "The relation between Granger causality and directed information theory: a review," *Entropy*, vol. 15, no. 1, pp. 113–143, Nov. 2012. [Online]. Available: http://dx.doi.org/10.3390/e15010113

[20] P. O. Amblard and O. J. J. Michel, "On directed information theory and Granger causality graphs," *Journal of Computational Neuroscience*, vol. 30, no. 1, pp. 7–16, Feb. 2010. [Online]. Available: http://dx.doi.org/10.1007/s10827-010-0231-x

[21] N. Ancona, D. Marinazzo, and S. Stramaglia, "Radial basis function approach to nonlinear Granger causality of time series," *Phys. Rev. E*, vol. 70, pp. 056 221+, Nov. 2004. [Online]. Available: http://dx.doi.org/10.1103/physreve.70.056221

[22] A. M. Bastos and J.-M. M. Schoffelen, "A Tutorial Review of Functional Connectivity Analysis Methods and Their Interpretational Pitfalls." *Frontiers in systems neuroscience*, vol. 9, 2015. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/26778976

[23] C. W. J. Granger, "Testing for causality," *Journal of Economic Dynamics and Control*, vol. 2, no. 2-4, pp. 329–352, Jan. 1980. [Online]. Available: http://dx.doi.org/10.1016/0165-1889(80)90069-x

[24] 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, pp. 238 701+, Nov. 2009. [Online]. Available: http://dx.doi.org/10.1103/physrevlett.103.238701

[25] 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, Jun. 1982. [Online]. Available: http://dx.doi.org/10.1080/01621459.1982.10477803

[26] D. Chicharro, "On the spectral formulation of Granger causality," *Biological Cybernetics*, vol. 105, no. 5-6, pp. 331–347, Dec. 2011. [Online]. Available: http://dx.doi.org/10.1007/s00422-011-0469-z

[27] 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, Jun. 2001. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/11417058

[28] M. J. Kaminski and K. J. Blinowska, "A new method of the description of the information flow in the brain structures," *Biological Cybernetics*, vol. 65, no. 3, pp. 203–210, Jul. 1991. [Online]. Available: http://dx.doi.org/10.1007/bf00198091

[29] D. Y. Takahashi, L. A. Baccala, and K. Sameshima, "Frequency domain connectivity: an information theoretic perspective." *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, vol. 2010, pp. 1726–1729, 2010. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/21096407

[30] E. Pereda, R. Quian, and J. Bhattacharya, "Nonlinear multivariate analysis of neurophysiological signals." *Progress in neurobiology*, vol. 77, no. 1-2, pp. 1–37, Sep. 2005. [Online]. Available: http://dx.doi.org/10.1016/j.pneurobio.2005.10.003

[31] D. Y. Takahashi, L. A. Baccalá, and K. Sameshima, "Information theoretic interpretation of frequency domain connectivity measures," Dec. 2010. [Online]. Available: http://arxiv.org/abs/1012.0353

[32] A. B. Barrett, L. Barnett, and A. K. Seth, "Multivariate Granger Causality and Generalized Variance," *Physical Review E*, vol. 81, no. 4, pp. 041 907+, Apr. 2010. [Online]. Available: http://dx.doi.org/10.1103/physreve.81.041907

[33] 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, Dec. 1984. [Online]. Available: http://dx.doi.org/10.1080/01621459.1984.10477110

[34] 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, Feb. 2014. [Online]. Available: http://dx.doi.org/10.1016/j.jneumeth.2013.10.018

[35] M. Vinck, L. Huurdeman, C. A. Bosman, P. Fries, F. P. Battaglia, C. M. Pennartz, and P. H. Tiesinga, "How to detect the Granger-causal flow direction in the presence of additive noise?" *NeuroImage*, vol. 108, pp. 301–318, Mar. 2015. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/25514516

[36] K. Mizuseki, A. Sirota, E. Pastalkova, and G. Buzsáki, "Theta oscillations provide temporal windows for local circuit computation in the entorhinal-hippocampal loop." *Neuron*, vol. 64, no. 2, pp. 267–280, Oct. 2009. [Online]. Available: http://dx.doi.org/10.1016/j.neuron.2009.08.037

[37] K. Mizuseki, A. Sirota, E. Pastalkova, K. Diba, and G. Buzsáki, "Multiple single unit recordings from different rat hippocampal and entorhinal regions while the animals were performing multiple behavioral tasks. CRCNS.org." 2013.

[38] N. M. van Strien, N. L. M. Cappaert, and M. P. Witter, "The anatomy of memory: an interactive overview of the parahippocampalhippocampal network," *Nature Reviews Neuroscience*, vol. 10, no. 4, pp. 272–282, Apr. 2009. [Online]. Available: http://dx.doi.org/10.1038/nrn2614

[39] P. Domingos, "A Few Useful Things to Know About Machine Learning," *Commun. ACM*, vol. 55, no. 10, pp. 78–87, Oct. 2012. [Online]. Available: http://dx.doi.org/10.1145/2347736.2347755

Chapter 9

Validating Unsupervised and Supervised Brain Connectivity Inference Methods with Realistic Neural Network Simulations

Validating Unsupervised and Supervised Brain Connectivity Inference Methods with Realistic Neural Network Simulations

Danilo Benozzo, Jan Bin, Stefano Panzeri, Paolo Avesani.

***Abstract*—The study of causality is of high interest in neuroscience since it allows the understanding of the network of direct interactions between brain areas.**

Causality is typically studied from observational data and, in order to facilitate the interpretation of the result from the neuroscience point of view, a parametric approach should be preferred.

As parametric approach, we refer to a causal estimate based on a parametric implementation of a certain criterion of causality. The criterion that we consider in this work is the Granger one and, as its implementation, we consider the multivariate autoregressive.

The purpose of this paper is to evaluate an unsupervised and a supervised brain connectivity inference methods when applied to a realistic neural network simulation. A major difference between unsupervised and supervised methods is that in the unsupervised ones the step of model identification is required to infer causal interactions while in the supervised one causal interactions are estimated after a learning phase involving a dedicated process of data generation.

In this work, we investigate the implications of having in the supervised approach a training phase with its own stochastic process. Specifically, we exploit the possibility of defining the stochastic process of the training phase equals to the neural network model used to generate the realistic simulations. And we conclude that this leads to a more accurate inference and makes the approach more application dependent avoiding the issue of model identification.

I. Introduction

In neuroscience, communication among neurons or among brain regions is typically studied from observational data through the so-called causal connectivity analysis [1].

Causal connectivity refers to a statistical causality measure that infers the direct dynamical interactions among neurons from the observation of their activity. Only in the specific case of an inference process that physiologically models both the structure and the dynamics of the neuronal activity, the causality measure can be interpreted in neuroscientific terms to infer elements of a neural circuit diagram. Thus, this causality measure can be considered as an estimate of the effective connectivity [2].

The need to have a generative model in order to facilitate the interpretation of the estimate, implies the choice of a parametric approach for the causal inference. As parametric approach, we refer to a causal estimate based on a model that encodes a certain criterion of causality and eventually also features of the physiological process.

Regarding the parametric approach of causal inference, the multivariate autoregressive (MAR) formulation of the Granger criterion of causality is one of the most studied in the literature. This modelling of the process does not consider the physiological mechanism of generation and interaction but simply models the dependence between past and present time points among time series. Indeed, the MAR model is not constrained to a specific type of signals as suggested by its wide application in many fields of science.

The MAR model implies to assume to directly measure the processes between which the causal interactions are exhibited. From the neuroscience point of view, this may be a very strong assumption. To this purpose, in [3] the so-called partial Granger causality has been introduced in order to deal with the problem of hidden (unrecorded) variables.

In addition to this example, many other measures of causality have been developed from the MAR formulation of the Granger criterion. We just mention the Geweke measure in the time domain under the assumption of stationary stochastic processes [4]. The Geweke measure is an example of conditional pair wise measure and it evaluates whether the past of the candidate cause significantly improves the prediction accuracy of the future of the candidate effect. The comparison is done in terms of variances of the residuals of two MAR models that differ in the presence or absence of the candidate cause.

To summarize, our main comment on the MAR implementation of the Granger criterion concerns the model itself. On one hand, the model identification is straightforward, it is done through standard linear autoregressive methods e.g. ordinary least square and related variations to impose sparsity constrains, multivariate Yule-Walker equations etc. [1]. On the other hand, this simple model suffers for lack of realism in several applications. Moreover, beyond the simplicity, the MAR model also assumes that the observed time series are a direct measurement of the neural processes. This is a realistic approximation only in the case of invasive recordings, but not in general [5].

The scope of this paper is to investigate the effect of changing the generative model when causality is inferred by a parametric method. The common practice is to evaluate the predictive capability of a causal measure by testing it on the

D. Benozzo and P. Avesani are with the NeuroInformatics Laboratory (NILab), Bruno Kessler Foundation, Trento, Italy and with the Center for Mind and Brain Sciences (CIMeC), University of Trento, Italy.

J. Bin and S. Panzeri are with the Neural Computation Laboratory, Italian Institute of Technology Center for Neuroscience and Cognitive Systems @ UniTn, Trento, Italy.

parametric model from which it is derived, under different levels and types of noise [6]. Here, we want to break the consistency between the generative model and the parametric implementation of the causal criterion and we want to study how the inference reacts to it. We take advantage of the fact that recent progress in neural network modelling makes it possible to generate models of recurrent microcircuits that have biophysical and anatomical properties very similar to those of real cortical circuits and that, importantly, when simulated as dynamical systems, generate activity with statistics very close to that of recorded cortical activity [7]–[10]. By connecting several of these simulated microcircuits we can, as shown here, generate a simulated information flow among neural circuits that has realistic statistical properties and for which we know the ground truth of causal communication. The study aims to take advantage of the availability of these simulated data and to show how the performance of the parametric causal inference changes when it is approached both in an unsupervised and a supervised way, and a physiological plausible model is used as generative model. Regarding the parametric unsupervised method, we chose the Geweke measure in time and for its inference we used the toolbox proposed in [11]. As parametric supervised method we refer to the technique proposed in [12]. In short, the idea of the supervised method is to unveil causal connections given a set of signals, through a classification schema. This needs a training dataset that is generated by the adopted generative model and a feature space which is defined according to a predefined criterion of causality. These two approaches are analysed under two causal models. Firstly, the MAR model is assumed, thus in line with the assumptions of the Geweke measure. In the second case, the physiological neuronal model is adopted in which multiple neuronal models are connected together in order to built a network of a given structure.

II. Materials and Methods

In this Section, we provide details on the datasets generation and on the parametric methods used for the causal inference.

A. MAR model and dataset

The multivariate autoregressive (MAR) dataset contains multiple trials and each trial $\mathbf{X} = \{X(t), t = 0, 1, \ldots, N-1\}, X(t) \in \mathbb{R}^{M\times 1}$ is defined as the linear combination of two M-dimensional multivariate time series $\mathbf{X_s}$ and $\mathbf{X_n}$

$$\mathbf{X} = (1-\gamma)\mathbf{X_s} + \gamma\mathbf{X_n} \tag{1}$$

$\mathbf{X_s}$ carries the causal information, $\mathbf{X_n}$ represents an additive noise corruption and $\gamma \in [0, 1]$ tunes the signal-to-noise ratio. Each time point of $\mathbf{X_s}$ and $\mathbf{X_n}$ is computed by following the stationary MAR model

$$\begin{aligned} X_s(t) &= \sum_{\tau=1}^{p} A_s^{(\tau)\top} X_s(t-\tau) + \varepsilon_s(t) \\ X_n(t) &= \sum_{\tau=1}^{p} A_n^{(\tau)\top} X_n(t-\tau) + \varepsilon_n(t) \end{aligned} \tag{2}$$

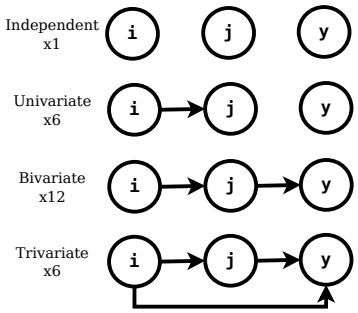


Fig. 1: Possible DAGs exhibited among 3 labeled nodes used in the dataset generation.

where p is the order of the MAR model and represents the maximal time lag. $\varepsilon_s(t)$ and $\varepsilon_n(t)$ are the innovation processes, defined as realizations from a diagonal M-dimensional standard normal distribution. $A_s^{(\tau)}, A_n^{(\tau)} \in \mathbb{R}^{M\times M}, \tau = 1, \ldots, p$ are the coefficient matrices modelling the influence of the signal values at time $t-\tau$ on the current signal values, i.e. at time t. The coefficient matrices $A_s^{(\tau)}$ defines the process of causal-informative data generation. They are computed by randomly corrupting the non-zero elements of the $M \times M$ binary matrix A, called causal configuration matrix. In essence, A represents the causal graph that leads the MAR model. Specifically $A_{ij} = 1$ means signal i causes the signal j. On the other hand, coefficient matrices $A_n^{(\tau)}$ lead the noisy part of the signals and they are obtained by randomly generating p diagonal matrices. In our experimental setup, we chose $M = 3$, $p = 10$ and $N = 6000$. Moreover, for each causal configuration matrix, 1000 trials were generated. From now on we will refer to this dataset as L dataset. Regarding the causal configurations, we considered the possible DAGs exhibited among 3 labeled nodes, see Figure 1. Equivalence between DAGs was not taken into account thus the total number of configurations is 25.

B. Neural Network (NN) model and dataset

In order to generate data qualitatively as similar as possible to physiological recordings, we decided to use a model based on work of Mazzoni et. al [8]. In their work, they presented a model of cortical network composed of leaky integrate and fire neurons and they managed to obtain behavior that strongly resembles primary visual cortex. For our purposes, we used the model separately for three populations of neurons that were connected with each other based on intended scenario.

The simulated network is composed of $N = 5000$ neurons. 80% of the neurons are taken to be excitatory, the remaining 20% are inhibitory [13]. The network is randomly connected: the connection probability between any directed pair of cells is 0.2 [14], [15]. In case of an inter-network directed connection, there is also 0.2 probability of connection between any pair composed of any cell from the receiver network and an excitatory cell from the sender network, Fig. 2. Both pyramidal (excitatory) neurons and interneurons (inhibitory) are described by leaky integrate and fire (LIF) dynamics [16].

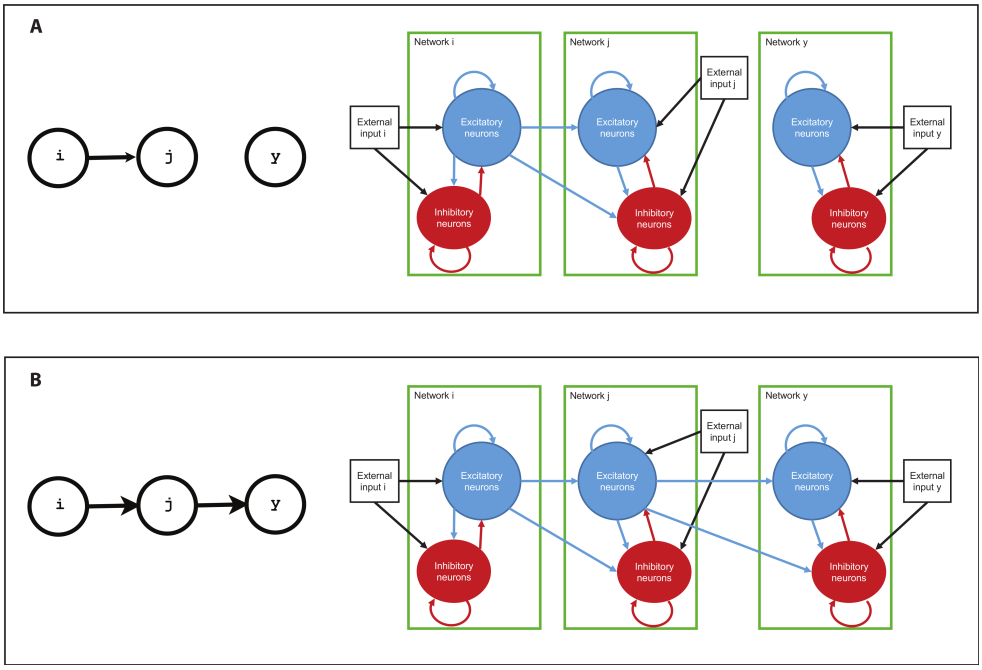


Fig. 2: The connectivity of the network; Both populations of neurons, excitatory and inhibitory have connections within the population and also with each other. External input is connected to every neuron in a given network. Finally, if a connection between networks is present, there are directed connections from excitatory population of the sender to both populations in the receiving network. **A** Represents univariate connection from network i to network j. **B** Depicts bivariate connection between networks. The information flows from network i to network j and from there to network y.

Each neuron k is described by its membrane potential V_k that evolves according to

$$\tau_m \frac{dV_k}{dt} = -V_k + I_{Ak} - I_{Gk} \tag{3}$$

where τ_m is the membrane time constant (20 ms for excitatory neurons, 10 ms for inhibitory neurons, [17]), I_{Ak} are the (AMPAtype) excitatory synaptic currents received by neuron k, while I_{Gk} are the (GABA-type) inhibitory currents received by neuron k. Note that in (3) we have taken the resting potential to be equal to zero. When the membrane potential crosses the threshold V_{thr} (18 mV above resting potential) the neuron fires, causing the following consequences: i) the neuron potential is reset at a value V_{res} (11 mV above resting potential), ii) the neuron can not fire again for a refractory time τ_{rp} (2 ms for excitatory neurons, 1 ms for inhibitory neurons).

Synaptic currents are the linear sum of contributions induced by single pre-synaptic spikes, which are described by a difference of exponentials. They can be obtained using auxiliary variables x_{Ak}, x_{Gk}. AMPA and GABA-type currents of neuron k are described by

$$\tau_{dA} \frac{dI_{Ak}}{dt} = -I_{Ak} + x_{Ak} \tag{4}$$

$$\begin{aligned} \tau_{rA} \frac{dx_{Ak}}{dt} &= -x_{Ak} + \\ +\tau_m \Bigg(J_{k-exc} \sum_{exc} \delta(t - t_{k-exc} - \tau_L) &+ \\ + J_{k-int} \sum_{int} \delta(t - t_{k-int} - \tau_{L-int}) &+ \\ + J_{k-ext} \sum_{ext} \delta(t - t_{k-ext} - \tau_L) \Bigg) & \end{aligned} \tag{5}$$

$$\tau_{dG} \frac{dI_{Gk}}{dt} = -I_{Gk} + x_{Gk} \tag{6}$$

$$\begin{aligned} \tau_{rG} \frac{dx_{Gk}}{dt} &= -x_{Gk} + \\ +\tau_m \left(J_{k-inh} \sum_{inh} \delta(t - t_{k-inh} - \tau_L) \right) & \end{aligned} \tag{7}$$

where $t_{k-exc/inh/int/ext}$ is the time of the spikes received from excitatory neurons/inhibitory neurons/inter-network exc. neurons (if a connection from another network is present) connected to neuron k, or from external inputs (see below). τ_{dA} (τ_{dG}) and τ_{rA} (τ_{rG}) are respectively the decay and rise time of the AMPA-type (GABA-type) synaptic current. $\tau_L = 1ms$ and $\tau_{L-int} = 3$ ms are latencies of post-synaptic currents for intra- and inter-network connections respectively.

$J_{k-exc/inh/int/ext}$ is the efficacy of the connections from excitatory neurons/inhibitory neurons/inter-network exc. neurons/external inputs on the population of neurons to which k belongs.

Each neuron is receiving an external excitatory synaptic input (last term in the r.h.s. of (5)). These synapses are activated by random Poisson spike trains, with a time varying rate which is identical for all neurons. This rate is given by

$$\nu_{ext}(t) = [\nu_{signal}(t) + n(t)]_+ \tag{8}$$

where $\nu_{signal}(t)$ represents the signal, and $n(t)$ is the noise. $[\cdots]_+$ is a threshold-linear function, $[x]_+ = x$ if $x > 0$, $[x]_+ = 0$ otherwise, to avoid negative rates which could arise due to the noise term. We use constant signal defined by

$$\nu_{signal}(t) = \nu_0 \tag{9}$$

where ν_0 is a constant rate equal to 2 spikes/ms. The noise represented by $n(t)$ in (8) is generated according to an Ornstein-Uhlenbeck process.

The activity of each network was summarized by generation of simulated local field potential (LFP). To capture in a simple way the fact that pyramidal cells contribute the most to LFP generation the LFPs are modeled as the sum of the absolute values of AMPA and GABA currents ($|I_A| + |I_G|$) on pyramidal cells in every time point of the simulation.

In all scenarios of Fig. 1 we simulated three networks with the same set of parameters. However, their internal connections and external inputs were generated independently. All the parameter values were in agreement with the original work of Mazzoni [8] with addition of synaptic efficacies for inter-network connections J_{k-int} that were equal for excitatory and inhibitory neurons and were drawn from a uniform distribution from interval $\langle 0, 0.18)$ for every pair of networks in every trial. From now on will refer to this dataset as NN dataset.

C. Parametric methods for causal inference

Both the two methods for causal inference that are used in this paper, assume the MAR model in their parameterization. In particular, the Geweke measure is a linear measure of Granger causality and it is based on the MAR process theory.

Consider a system of three stationary stochastic processes X_t, Y_t and Z_t. The pair-wise conditional approach examines whether Y has a direct influence on X given the presence of Z by decomposing

$$X_t = \sum_{i=1}^{\infty} a_{xx,i} X_{t-i} + \sum_{i=1}^{\infty} a_{xy,i} Y_{t-i} + \sum_{i=1}^{\infty} a_{xz,i} Z_{t-i} + \varepsilon_{x,t} \tag{10}$$

Afterwards, the reduced autoregressive representation of X is considered

$$X_t = \sum_{i=1}^{\infty} a'_{xx,i} X_{t-i} + \sum_{i=1}^{\infty} a'_{xz,i} Z_{t-i} + \varepsilon'_{x,t} \tag{11}$$

The Geweke index of causality in time domain $F_{Y \to X|Z}$ evaluates which of the two regressions (10) and (11) models better the process X by computing

$$F_{Y \to X|Z} = \ln \frac{\Sigma'_{xx}}{\Sigma_{xx}} \tag{12}$$

where $\Sigma'_{xx} = \text{var}(\varepsilon'_{xx})$ and $\Sigma_{xx} = \text{var}(\varepsilon_{xx})$ are the residual variances of the MAR models (10) and (11) respectively. (12) is interpreted as the variation in prediction error when the past of Y is included in the regression. An important aspect is the statistical significance of the estimated causal measure and the common practice is to look at (12) as the test statistic of a log-likelihood ratio test. In particular, it results that under the null hypothesis of zero causality $H_0 : a_{xy,i} = 0, \forall i$ the Geweke measure has an asymptotic χ^2 distribution up to a scaling factor which depends on the sample size and with degree of freedom equals to the difference in the number of parameters between (10) and (11) models. Under the alternative hypothesis, the scaled test statistic has an asymptotic noncentral χ^2 distribution with noncentrality parameter that corresponds to the scaled casual measure. In a more general formulation the three processes may be multivariate thus they may represent a set of variables. In our experiments, we used the implementation proposed in [11] and we will refer to it as the Granger Causal Analysis method (GCA). It is important to underline the unsupervised nature of this method. This will be crucial in understanding the result part and related discussion. In order words, we refer to the fact that given a certain criterion of causality a measure of causality is derived by a parametric implementation of this criterion. By implementing in a parametric way a certain criterion, a generative model is implicitly assumed, thus also the working assumption of the measure and its best working scenario consequently derive.

Regarding the supervised parametric approach of which we refer to as SL, it frames the causal inference as a learning theory problem. A model of the stochastic process is used for generating a representative dataset of the population of causal graphs of interest instead of implementing a certain criterion as in the unsupervised case. This dataset is used to train a classifier which will be applied to predict the causal graph of an unseen trial of time series. A key point of this approach is the definition of the feature space that is expected to emphasize the causal structure of the trial in order to facilitate the inference. A detailed description of the feature space is given in [12]. Here, it is relevant to stress that the feature space is defined on the autoregressive implementation of the Granger criterion while the stochastic process is described by the generative model used for the training set generation. This allows the generative model and the implementation of the causal criterion to be treated separately. And thanks to the training phase, the chosen causal criterion is shaped on the adopted generative model.

III. Experiments and Results

In this section, experiments are described. In particular, they are grouped in two parts. The first part focuses on the NN dataset and it aims to investigate the generated trials in terms of how the direct connection is exhibited. While, in the second part the two methods for causal inference are applied on the two datasets.

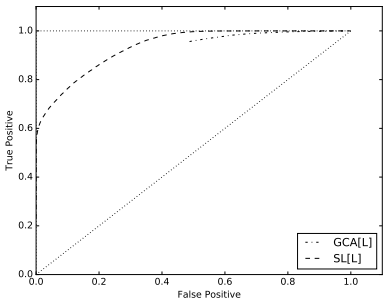


Fig. 5: ROC curves from the application of GCA and SL in the L dataset.

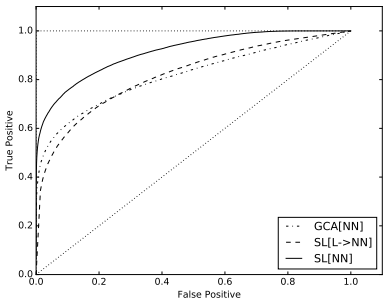


Fig. 6: ROC curves from the application of GCA and SL in the NN dataset. SL is applied twice with different training phases. SL[L→NN] indicates that the method was trained on L while for SL[NN] the training was done directly on the NN dataset.

A. Characterization of the NN signal

a) Cross-correlation: We computed cross-correlation of each pair of activities for all networks and report the results for two causal configurations in Figure 3. In the other configurations we observe the same behavior. The cross-correlation confirms connections between the networks, yielding the peak shifted from 0, where it would be expected based on the autocorrelation of the signal, to the time corresponding to the transfer delay (3ms). Moreover, it captures the transitivity of those connections, showing a peak shift between networks that are not connected directly but via an another one.

b) Geweke measure in time domain: GCA was applied in each trial of the causal configurations that are reported in Figure 4 and the related results are shown by plotting the log p-value distribution of each ordered pair of nodes. We can see from the figure that if the nodes are not connected then the p-values are approximately uniformly distributed, i.e. the null hypothesis is accepted. While in the case of connection, there is an higher concentration of low p-values under the two vertical dashed thresholds that correspond to 0.01 and 0.05.

B. Effect of the generative model on parametric methods

As preliminary experiment, we inferred the causality in L by applying the GCA toolbox and the supervised approach (SL), Figure 5 and Table I. This first experiment is meant to compare the two approaches in a scenario in which their working assumptions completely hold. Indeed, in both cases the generative model (MAR) is coherent with the implementation of the causal criterion in the methods. The only component that makes the inference challenging is the signal-to-noise ratio since the γ parameter is uniformly distributed in $[0, 1]$, see Subsection II-A. The related outcomes highlight the issue of GCA of overestimating the connections in the presence of additive noise.

	GCA[L]	SL[L]
AUC	0.7409±0.0011	0.9350±0.0003

TABLE I: AUC values related to the application of GCA and SL in the L dataset.

The second experiment repeats the same analysis on the NN dataset. Results are shown in Figure 6 and Table II and are named: GCA[NN] and SL[L→NN]. Both methods assume the MAR model as generative model while they are applied on trials generated by the NN model. Moreover, the NN model does not implement a casual link by directly following the MAR implementation. Even though, as we saw in Figure 4, the Granger criterion seems to be appropriate to infer the causal connections of NN.

Regarding the third experiment, its purpose is to test whether the inference improves by keeping the generative model of the training dataset consistent with the generative model of the dataset in which the method will be applied. Thus, we refer to SL[NN] in the Figure 6 and Table II. SL was applied to the NN dataset after having been trained on the same model and by using the Granger-based feature space.

	GCA[NN]	SL[L→NN]	SL[NN]
AUC	0.8160±0.0008	0.8178±0.0007	0.9139±0.0005

TABLE II: AUC values related to the application of GCA and SL on the NN dataset.

IV. Discussion

The purpose of this research activity was to investigate the effect of changing the generative model when causality is inferred by a parametric method. We considered two parametric approaches for time series causality: the standard autoregressive implementation of the Granger criterion in the time domain (GCA) and the supervised method (SL) based on the same criterion. Since both GCA and SL are parametric methods, they assume a generative model for the stochastic process of data generation. Regarding GCA, it is the MAR model while in the case of SL we should distinguish between the criterion adopted in the feature space and the generative model of the training set. Specifically, the feature space is defined by considering the MAR implementation of the Granger criterion. And about the generation of the training set, two alternatives are considered: the MAR model and the NN model.

The fact that SL allows the generation of the training set to be separated by the criterion used in the feature space, is

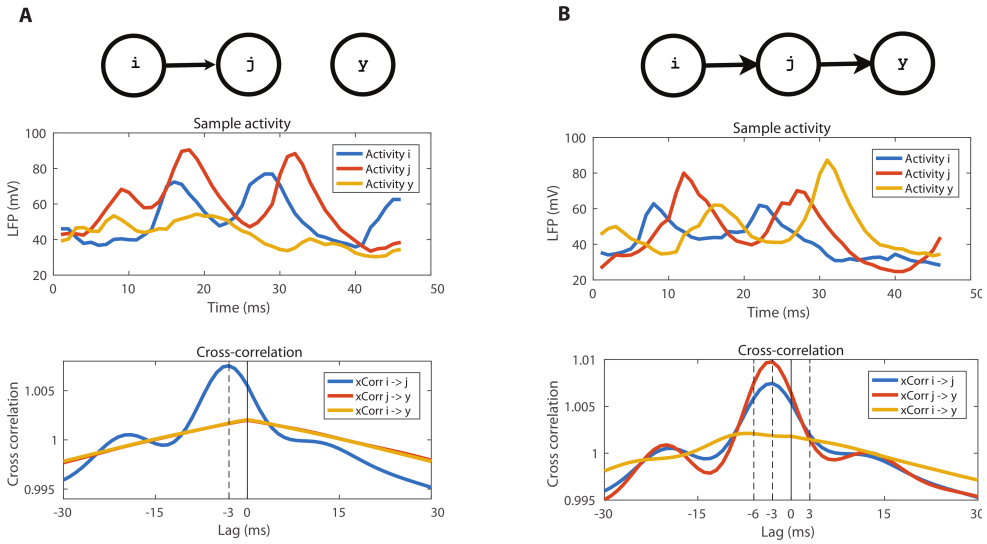


Fig. 3: Properties of the data generated by the NN model. The top panel described the architecture of the network, the middle one shows a sample of the activity of those networks and finally the last, cross-correlation between the activity.**A** In case of the univariate connection, it can be observed that activity of j follows activity of i, which is also confirmed by the cross-correlation peak at the time equal to the transfer delay. **B** Also in the case of bivariate transfer, it can be observed that the activities follow each other accordingly to the architecture of the network and that it is also confirmed by the cross-correlation. The peak of cross-correlation between i and y also shifted because of the transitivity of those connections.

the core aspect of this work. To this purpose, we evaluate three different scenarios: i) GCA and SL were applied in the L dataset, about SL the MAR model was used for the training set, ii) GCA and SL were applied in the NN dataset, and SL was trained as before on the MAR model, iii) SL was applied in the NN dataset after having training it on a dataset generated from the same model.

We remark that scenarios (i) and (ii) do a comparison between methods when applied to the same dataset and we notice that a comparison across datasets (keeping the same method) would not be not fair because it would not explain the inference capability of the method. Indeed, the L and NN datasets are not generated in order to differ only in the encoding of the causal interaction, so other confounds would bias the comparison, e.g. the signal-to-noise ratio.

The choice of the NN model was motivated by considering the real application scenario, and the intrinsic bias due to the inability of having a fully realistic generative model. Thus the idea of choosing a generative model that is not strictly based on the MAR implementation but it is more neuro-physiologically plausible.

Firstly, we ensured that the expected causal connections generated by NN were detectable in terms of cross-correlation and Granger causality. These simple preliminary analyses show at the qualitative level that the dependencies that have emerged, follow the ground truth. In particular, from Figure 4 we can see that the distribution of the p-values of the Geweke test statistic is not uniform when the two nodes are directly connected, i.e. the null hypothesis of zero causality is not accepted. This visual inspection of the p-value distributions gives evidence that the interaction between nodes as it is modeled by NN, can be detected by the Granger criterion.

Regarding the second part of Section III and in particular in Figure 5, we can see the outcomes of the causal inference done on L by the two methods. Both methods are parametric and based on the MAR implementation of the Granger definition of causality. By applying them in L, their working assumptions hold both from the side of the data generation itself and the side of how the model encodes a causal interaction. The additive noise included in the model is the only interfering component in the inference process. Thus this experiment is mainly meant for evaluating the two approaches on the ideal case in which just the noise confounds the inference. And from Figure 5 and Table I we conclude that SL is more robust than GCA to the presence of noise.

The second experiment repeats the same analysis on the NN dataset. Differently from the previous case, the working hypotheses of both methods do not completely hold when applied in the NN dataset. In particular, in the case of GCA neither the generative assumption nor how a causal interaction is implemented, are satisfied when it is applied in NN. The same is true when SL is trained on the MAR dataset and applied in NN. This experiment emulates the empirical application of GCA, under the hypothesis that the NN dataset well reproduces a real neural recording. This hypothesis is motivated by the neuro-physiological plausibility of the NN model. Regarding the supervised approach, it was designed in order to run under the same condition of GCA. Indeed the training was done on the MAR dataset and the feature space was defined by considering the MAR implementation of the

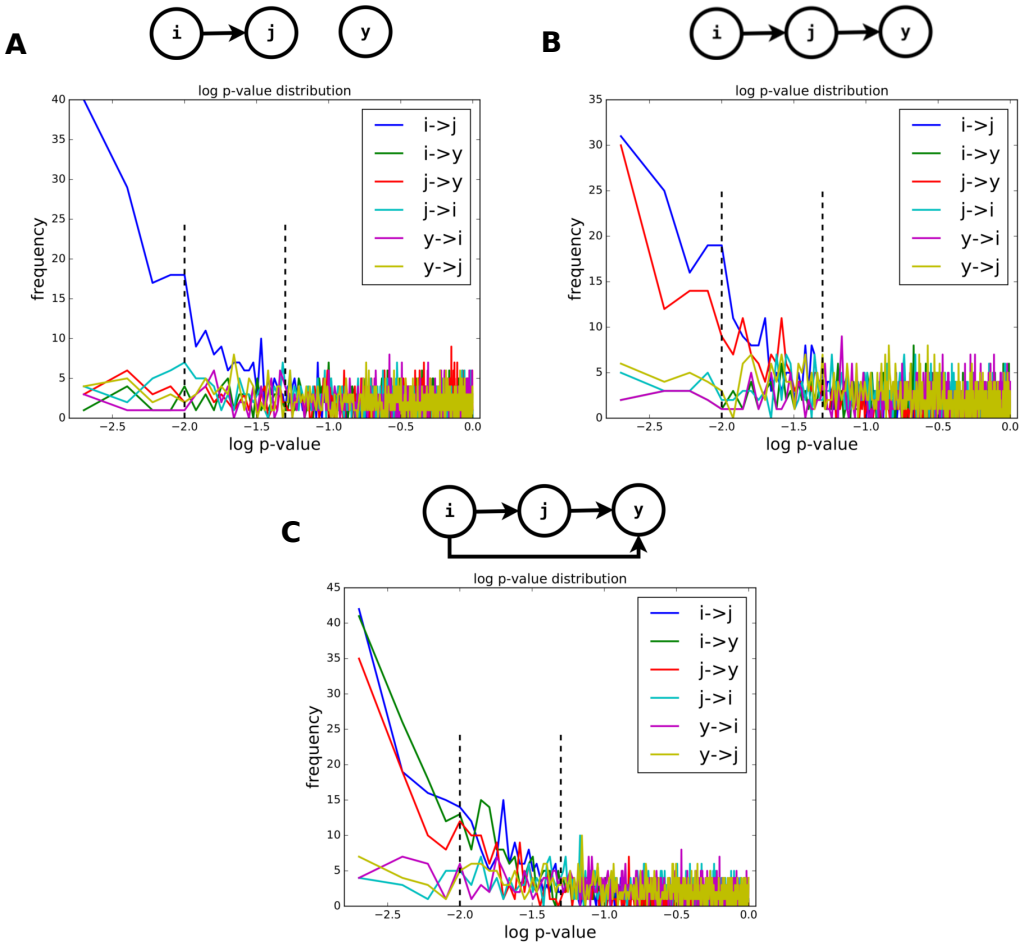


Fig. 4: The log p-value distributions of the Geweke test statistic computed by GCA for each ordered pair of nodes. Three classes of trials were selected respectively from the group of univariate **A**, bivariate **B** and trivariate **C** configurations.

Granger causality. Said differently, in both cases neither the actual nature of the data nor the actual encoding of a causal interaction, were considered. From the results that are reported in Figure 5 and Table II, it emerges that the performances of these analyses are very similar.

The last experiment investigates the supervised learning approach when the generative model does not derive from the causal criterion. To evaluate this scenario, we infer the causality in the NN dataset by training SL on a feature space whose features are Granger-based but it is constructed on a dataset generated by the NN model. By constraining the training phase by assuming the same generative model of the testing dataset, the accuracy of detecting the correct causal interactions substantially improves. In Figure 6, the improvement is shown by the solid line and quantified in Table II in which we see the difference with the accuracies of the previous experiment.

V. Conclusion

In this paper, we validated two different approaches for the inference of causal brain connectivity by using realistic neural network simulations. Specifically, our focus was on the inference performances of the unsupervised and supervised methods when the assumptions of data generation do not hold. Moreover, we investigated the implications of having in the supervised approach a training phase with its own stochastic process which does not directly depend on the causal criterion that instead it is implemented in the feature space definition. This aspect of the supervised approach leads to a more accurate inference and makes it more application dependent without the problem of model identification. As future work, we plan to extend the same analysis also on a real dataset.

References

[1] S. L. Bressler and A. K. Seth, "Wiener-Granger causality: a well established methodology." *NeuroImage*, vol. 58, no. 2, pp. 323–329,

Sep. 2011. [Online]. Available: http://dx.doi.org/10.1016/j.neuroimage.2010.02.059

[2] K. J. Friston, "Functional and effective connectivity: a review." *Brain Connectivity*, vol. 1, no. 1, pp. 13–36, 2011. [Online]. Available: http://dx.doi.org/10.1089/brain.2011.0008

[3] S. Guo, A. K. Seth, K. M. Kendrick, C. Zhou, and J. Feng, "Partial Granger causality–eliminating exogenous inputs and latent variables." *Journal of neuroscience methods*, vol. 172, no. 1, pp. 79–93, Jul. 2008. [Online]. Available: http://dx.doi.org/10.1016/j.jneumeth.2008.04.011

[4] D. Chicharro, "On the spectral formulation of Granger causality," *Biological Cybernetics*, vol. 105, no. 5-6, pp. 331–347, Dec. 2011. [Online]. Available: http://dx.doi.org/10.1007/s00422-011-0469-z

[5] P. A. Valdes-Sosa, A. Roebroeck, J. Daunizeau, and K. Friston, "Effective connectivity: influence, causality and biophysical modeling." *NeuroImage*, vol. 58, no. 2, pp. 339–361, Sep. 2011. [Online]. Available: http://dx.doi.org/10.1016/j.neuroimage.2011.03.058

[6] M. Vinck, L. Huurdeman, C. A. Bosman, P. Fries, F. P. Battaglia, C. M. Pennartz, and P. H. Tiesinga, "How to detect the Granger-causal flow direction in the presence of additive noise?" *NeuroImage*, vol. 108, pp. 301–318, Mar. 2015. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/25514516

[7] N. Brunel and X.-J. Wang, "What Determines the Frequency of Fast Network Oscillations With Irregular Neural Discharges? I. Synaptic Dynamics and Excitation-Inhibition Balance," *Journal of Neurophysiology*, vol. 90, no. 1, pp. 415–430, Jul. 2003. [Online]. Available: http://dx.doi.org/10.1152/jn.01095.2002

[8] A. Mazzoni, S. Panzeri, N. K. Logothetis, and N. Brunel, "Encoding of Naturalistic Stimuli by Local Field Potential Spectra in Networks of Excitatory and Inhibitory Neurons," *PLOS Computational Biology*, vol. 4, no. 12, pp. e1 000 239+, Dec. 2008. [Online]. Available: http://dx.doi.org/10.1371/journal.pcbi.1000239

[9] A. Mazzoni, K. Whittingstall, N. Brunel, N. K. Logothetis, and S. Panzeri, "Understanding the relationships between spike rate and delta/gamma frequency bands of LFPs and EEGs using a local cortical network model." *NeuroImage*, vol. 52, no. 3, pp. 956–972, Sep. 2010. [Online]. Available: http://dx.doi.org/10.1016/j.neuroimage.2009.12.040

[10] A. Mazzoni, H. Lindén, H. Cuntz, A. Lansner, S. Panzeri, and G. T. Einevoll, "Computing the Local Field Potential (LFP) from Integrate-and-Fire Network Models," *PLOS Computational Biology*, vol. 11, no. 12, pp. e1 004 584+, Dec. 2015. [Online]. Available: http://dx.doi.org/10.1371/journal.pcbi.1004584

[11] 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, Feb. 2014. [Online]. Available: http://dx.doi.org/10.1016/j.jneumeth.2013.10.018

[12] D. Benozzo, E. Olivetti, and P. Avesani, "Classification-Based Causality Detection in Time Series," in *Machine Learning and Interpretation in Neuroimaging*, ser. Lecture Notes in Computer Science, I. Rish, G. Langs, L. Wehbe, G. Cecchi, K.-m. K. Chang, and B. Murphy, Eds. Springer International Publishing, 2016, vol. 9444, pp. 85–93. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-45174-9_9

[13] V. Braitenberg and A. Schüz, *Anatomy of the cortex: Statistics and geometry*. Springer-Verlag Publishing, 1991.

[14] P. J. Sjöström, G. G. Turrigiano, and S. B. Nelson, "Rate, timing, and cooperativity jointly determine cortical synaptic plasticity," *Neuron*, vol. 32, no. 6, pp. 1149–1164, 2001.

[15] C. Holmgren, T. Harkany, B. Svennenfors, and Y. Zilberter, "Pyramidal cell communication within local networks in layer 2/3 of rat neocortex," *The Journal of physiology*, vol. 551, no. 1, pp. 139–153, 2003.

[16] H. C. Tuckwell, *Introduction to Theoretical Neurobiology: Volume 1, Linear Cable Theory and Dendritic Structure*. Cambridge University Press, 1988, vol. 1.

[17] D. A. McCormick, B. W. Connors, J. W. Lighthall, and D. A. Prince, "Comparative electrophysiology of pyramidal and sparsely spiny stellate neurons of the neocortex," *Journal of neurophysiology*, vol. 54, no. 4, pp. 782–806, 1985.

Bibliography

[1] Odd O. Aalen, Kjetil Røysland, Jon Michael M. Gran, and Bruno Ledergerber. Causality, mediation and time: a dynamic viewpoint. *Journal of the Royal Statistical Society. Series A, (Statistics in Society)*, 175(4):831–861, October 2012.

[2] P. O. Amblard and O. J. J. Michel. On directed information theory and Granger causality graphs. *Journal of Computational Neuroscience*, 30(1):7–16, February 2010.

[3] Pierre-Olivier Amblard and Olivier J. J. Michel. The relation between Granger causality and directed information theory: a review. *Entropy*, 15(1):113–143, November 2012.

[4] Nicola Ancona, Daniele Marinazzo, and Sebastiano Stramaglia. Radial basis function approach to nonlinear Granger causality of time series. *Phys. Rev. E*, 70:056221+, November 2004.

[5] L. A. Baccalá and K. Sameshima. Partial directed coherence: a new concept in neural structure determination. *Biological cybernetics*, 84(6):463–474, June 2001.

[6] Lionel Barnett and Anil K. Seth. The MVGC multivariate Granger causality toolbox: A new approach to Granger-causal inference. *Journal of Neuroscience Methods*, 223:50–68, February 2014.

[7] Adam B. Barrett, Lionel Barnett, and Anil K. Seth. Multivariate Granger Causality and Generalized Variance. *Physical Review E*, 81(4):041907+, April 2010.

[8] M. S. Bartlett. Some Aspects of the Time-Correlation Problem in Regard to Tests of Significance. *Journal of the Royal Statistical Society*, 98:536–543, 1935.

[9] André M. Bastos and Jan-Mathijs M. Schoffelen. A Tutorial Review of Functional Connectivity Analysis Methods and Their Interpretational Pitfalls. *Frontiers in systems neuroscience*, 9, 2015.

[10] Steven L. Bressler and Anil K. Seth. Wiener-Granger causality: a well established methodology. *NeuroImage*, 58(2):323–329, September 2011.

[11] M. J. Brookes, M. W. Woolrich, and G. R. Barnes. Measuring functional connectivity in MEG: a multivariate approach insensitive to linear source leakage. *NeuroImage*, 63(2):910–920, November 2012.

[12] Clemens Brunner, Martin Billinger, Martin Seeber, Timothy R. Mullen, and Scott Makeig. Volume Conduction Influences Scalp-Based Connectivity Estimates. *Frontiers in Computational Neuroscience*, 10, November 2016.

[13] D. Chicharro. On the spectral formulation of Granger causality. *Biological Cybernetics*, 105(5-6):331–347, December 2011.

[14] Daniel Chicharro and Anders Ledberg. When Two Become One: The Limits of Causality Analysis of Brain Dynamics. *PLoS ONE*, 7(3):e32466+, March 2012.

[15] Gopikrishna Deshpande, K. Sathian, and Xiaoping Hu. Effect of hemodynamic variability on Granger causality analysis of fMRI. *NeuroImage*, 52(3):884–896, September 2010.

[16] Michael Eichler. A graphical approach for evaluating effective connectivity in neural systems. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 360(1457):953–967, May 2005.

[17] K. J. Friston. Functional and effective connectivity: a review. *Brain Connectivity*, 1(1):13–36, 2011.

[18] Karl J. Friston, André M. Bastos, Ashwini Oswal, Bernadette van Wijk, Craig Richter, and Vladimir Litvak. Granger causality revisited. *NeuroImage*, 101:796–808, November 2014.

[19] John Geweke. Measurement of Linear Dependence and Feedback between Multiple Time Series. *Journal of the American Statistical Association*, 77(378):304–313, June 1982.

[20] John F. Geweke. Measures of Conditional Linear Dependence and Feedback between Time Series. *Journal of the American Statistical Association*, 79(388):907–915, December 1984.

[21] John C. Gore. Principles and practice of functional MRI of the human brain. *Journal of Clinical Investigation*, 112(1):4–9, July 2003.

[22] C. W. J. Granger. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, 37(3):424–438, August 1969.

[23] C. W. J. Granger. Testing for causality. *Journal of Economic Dynamics and Control*, 2(2-4):329–352, January 1980.

[24] Stefan Haufe and Arne Ewald. A Simulation Framework for Benchmarking EEG-Based Brain Connectivity Estimation Methodologies. *Brain topography*, June 2016.

[25] Stefan Haufe, Ryota Tomioka, Guido Nolte, Klaus-Robert Muller, and Motoaki Kawanabe. Modeling Sparse Connectivity Between Underlying Brain Sources for EEG/MEG. *Biomedical Engineering, IEEE Transactions on*, 57(8):1954–1963, August 2010.

[26] Stephen W. Hawking. *The No Boundary Condition and the Arrow of Time*, pages 346–357. Cambridge University Press, 1994.

[27] Barry Horwitz. The elusive concept of brain connectivity. *NeuroImage*, 19(2 Pt 1):466–470, June 2003.

[28] M. J. Kaminski and K. J. Blinowska. A new method of the description of the information flow in the brain structures. *Biological Cybernetics*, 65(3):203–210, July 1991.

[29] David Lopez-Paz, Krikamol Muandet, Bernhard Schölkopf, and Ilya Tolstikhin. Towards a Learning Theory of Cause-Effect Inference, May 2015.

[30] Dan Lysne and Dag Tjostheim. Loss of Spectral Peaks in Autoregressive Spectral Estimation. *Biometrika*, 74(1):200+, March 1987.

[31] Alberto Mazzoni, Stefano Panzeri, Nikos K. Logothetis, and Nicolas Brunel. Encoding of Naturalistic Stimuli by Local Field Potential Spectra in Networks of Excitatory and Inhibitory Neurons. *PLOS Computational Biology*, 4(12):e1000239+, December 2008.

[32] James M. McCracken. Exploratory Causal Analysis with Time Series Data. *Synthesis Lectures on Data Mining and Knowledge Discovery*, 8(1):1–147, March 2016.

[33] Franz H. Messerli. Chocolate consumption, cognitive function, and Nobel laureates. *The New England journal of medicine*, 367(16):1562–1564, October 2012.

[34] R. J. Moran, K. E. Stephan, T. Seidenbecher, H-C C. Pape, R. J. Dolan, and K. J. Friston. Dynamic causal models of steady-state responses. *NeuroImage*, 44(3):796–811, February 2009.

[35] Yves Nievergelt. A tutorial history of least squares with applications to astronomy and geodesy. *Journal of Computational and Applied Mathematics*, 121(1-2):37–72, September 2000.

[36] Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3(0):96–146, 2009.

[37] Judea Pearl. An Introduction to Causal Inference. *The International Journal of Biostatistics*, 6(2), January 2010.

[38] Judea Pearl. Comment: Understanding Simpson's Paradox. *The American Statistician*, 68(1):8–13, January 2014.

[39] Ernesto Pereda, Rodrigo Quian, and Joydeep Bhattacharya. Nonlinear multivariate analysis of neurophysiological signals. *Progress in neurobiology*, 77(1-2):1–37, September 2005.

[40] Tarmo M. Pukkila. An improved estimation method for univariate autoregressive models. *Journal of Multivariate Analysis*, 27(2):422–433, November 1988.

[41] Graham E. Quinn, Chai H. Shin, Maureen G. Maguire, and Richard A. Stone. Myopia and ambient lighting at night. *Nature*, 399(6732):113–114, May 1999.

[42] Jakob Runge. *Detecting and quantifying causality from time series of complex systems*. PhD thesis, Humboldt-Universität zu Berlin, 2014.

[43] SewardB Rutkove. Introduction to Volume Conduction. In AndrewS Blum and SewardB Rutkove, editors, *The Clinical Neurophysiology Primer*, pages 43–53. Humana Press, 2007.

[44] V. Sakkalis. Review of advanced techniques for the estimation of brain connectivity measured with EEG/MEG. *Computers in Biology and Medicine*, 41(12):1110–1117, December 2011.

[45] Josè M. Sanchez-Bornot, Eduarso Martinez-Montes, Agustin Lage-Castellanos, Mayrim Vega-Hernandez, and Pedro A. Valdes-Sosa. Uncovering Sparse Brain Effective Connectivity: a Voxel-Based Approach Using Penalized Regression. *Statistica Sinica*, 18:1501–1518, 2008.

[46] Björn Schelter, Jens Timmer, and Michael Eichler. Assessing the strength of directed influences among neural signals using renormalized partial directed coherence. *Journal of neuroscience methods*, 179(1):121–130, April 2009.

[47] Marleen B. Schippers, Remco Renken, and Christian Keysers. The effect of intra- and inter-subject variability of hemodynamic responses on group level Granger causality analyses. *NeuroImage*, 57(1):22–36, July 2011.

[48] Jan-Mathijs M. Schoffelen and Joachim Gross. Source connectivity analysis with MEG and EEG. *Human brain mapping*, 30(6):1857–1865, June 2009.

[49] T. Schreiber. Measuring information transfer. *Physical review letters*, 85(2):461–464, July 2000.

[50] Anil K. Seth, Adam B. Barrett, and Lionel Barnett. Granger Causality Analysis in Neuroscience and Neuroimaging. *The Journal of Neuroscience*, 35(8):3293–3297, February 2015.

[51] Anil K. Seth, Paul Chorley, and Lionel C. Barnett. Granger causality analysis of fMRI BOLD signals is invariant to hemodynamic convolution but not downsampling. *NeuroImage*, 65:540–555, January 2013.

[52] J. J. C. Smart and Hans Reichenbach. The Direction of Time. *The Philosophical Quarterly*, 8(30):72+, January 1958.

[53] Elliott Sober. Venetian Sea Levels, British Bread Prices, and the Principle of the Common Cause. *The British Journal for the Philosophy of Science*, 52(2):331–346, June 2001.

[54] Victor Solo. On causality and mutual information. In *Decision and Control, 2008. CDC 2008. 47th IEEE Conference on*, pages 4939–4944. IEEE, December 2008.

[55] Daniel Y. Takahashi, Luiz A. Baccala, and Koichi Sameshima. Frequency domain connectivity: an information theoretic perspective. *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, 2010:1726–1729, 2010.

[56] Daniel Y. Takahashi, Luiz A. Baccalá, and Koichi Sameshima. Information theoretic interpretation of frequency domain connectivity measures, December 2010.

[57] Pedro A. Valdes-Sosa. Spatio-temporal autoregressive models defined over brain manifolds. *Neuroinformatics*, 2(2):239–250, 2004.

[58] Pedro A. Valdes-Sosa, Alard Roebroeck, Jean Daunizeau, and Karl Friston. Effective connectivity: influence, causality and biophysical modeling. *NeuroImage*, 58(2):339–361, September 2011.

[59] N. M. van Strien, N. L. M. Cappaert, and M. P. Witter. The anatomy of memory: an interactive overview of the parahippocampalhippocampal network. *Nature Reviews Neuroscience*, 10(4):272–282, April 2009.

[60] Raul Vicente, Michael Wibral, Michael Lindner, and Gordon Pipa. Transfer Entropy–a Model-free Measure of Effective Connectivity for the Neurosciences. *J. Comput. Neurosci.*, 30(1):45–67, February 2011.