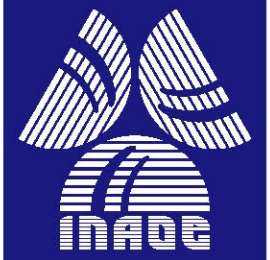


Learning Causal Probabilistic Graphical Models and their application to Effective Connectivity from functional Near InfraRed Spectroscopy

Samuel Antonio Montero Hernández

Technical Report No. CCC-16-009

National Institute for Astrophysics Optics and Electronics
Tonantzintla, Puebla, Mexico
June 28th 2016

INAOE, Mexico

Abstract

Science is, in general, concerned about determining causal relations between variables in order to give explanations about the past or future events. The area of causal discovery aggregates the set of theories, methods, and tools to recover causal models and make inferences from such models. This proposal is aimed at developing an approach for the causal discovery of probabilistic graphical models (PGMs). Considering a scenario where only a small set of samples is available in a system of interest, the addition of background information from the problem domain and the combination of interventional along with observational data, it is *hypothesized* that a causal PGM structure with greater number of invariant features can be recovered in contrast with the structure obtained if only that available set of observational data were considered, such that the true causal structure of the phenomenon of interest is more faithfully captured (i.e. maintaining, at least, the goodness of fit). Issues such as learning PGM structures with insufficient samples, relaxation of unmet assumptions problem domain and integration of different type of data (observational and experimental), will have to be addressed in order to develop such algorithm for learning causal structures based on PGMs in the afore-described conditions. Motivated by the problem of capturing causal relations in the human brain while a subject performs a specific task, functional near-infrared spectroscopy (fNIRS) neuroimaging will be used to interrogate neural activity and yield observational data with low cardinality imposed not only by experimenting but more physically limited by the inherent image formation physics and neurophysiological constraints. In a first experiment, the issue of incorporating background knowledge of the problem (structural information of the human connectome) have been addressed. An extension of the Fast Causal Inference (FCI) algorithm has been proposed to capture the brain connectivity network from a set of fNIRS neuroimages during an experiment for assessment of surgeons dexterity. Preliminary results confirm that the accuracy of the retrieved Causal PGM in terms of the number of resolved directed links improved in contrast with the accuracy of the Causal PGM retrieved when no background information was provided. As shown experimentally, the confirmatory results encourage to proceed with next stages of the methodology. At the same time, the use of the connectome information as complementary data motivates to investigate and encoding other neural circuits in order to find models of brain connectivity more accurate in terms of the directionality definition. The next stage of the work it is aimed to investigate how to make interventions "optimally" to generate additional data in order to improve the causal model. The main contributions of the thesis are (i) an algorithm for causal discovery overcoming or alleviating the classical requirement of a large set of sample, (ii) a mechanism to incorporate an efficient use of *a priori* information into a causal discovery algorithm, (iii) a method to select suitable target variables over which performing experimental disturbances plus to join the resulting interventional along with observational data aimed at elucidating the directionality of unresolved relations, and (iv) a causal PGM for the analysis of the

effective connectivity in fNIRS. With a contribution of an algorithm for causal discovery areas such as medicine, genetics, chemistry among others will be benefited by studying in more detail the relationships discovered by the proposed algorithm for causal discovery. Specifically, the field of neuroscience shall have a tool for understanding the function of the human brain in terms of the neural circuits using the fNIRS neuroimaging modality.

Keywords

causal discovery, probabilistic graphical models, effective connectivity, fNIRS

Contents

Chapter 1

Introduction

Science, in general, is concerned with the identification of relations between elements within a system. A special kind of them is the *cause-effect* relation. One way to explore the effect of a certain factor on dependent entities is by controlling all sources of variation and manipulating the factor of interest so that the differences in outcomes can be attributed to the manipulated variable. The interest in learning causal relations has permeated from exact to natural sciences. In particular, the study of brain function has received increased attention over the last decade (Friston, 2011), especially in the discovery of brain paths involved in sensorimotor or cognitive processes, the so-called brain connectivity analysis.

Frequently, the identification of causes cannot be established univocally because of several inherent limitations of the experiment deployed to understand the phenomenon under study. Limitations as insufficient measures, confounding factors or uncontrolled conditions may appear due to physical or ethical considerations, resulting in a failure to comply with the model's principal assumptions.

In causal discovery area, some issues hinder the application of current causal generation models because most of them need to count with large samples in order to learn the model's parameters, as well as assuming the ability to control every variable in the system during the experiment to ensure that reliable model is recovered. For this reason, it is desirable a framework capable of recovering the causal structure considering: a reduced set of observations, potential hidden causes, experimental/simulated perturbations and physical measurements under uncertainty. There is as far as we know, no PGM capable of addressing a scenario with all these limitations at once.

Probabilistic Graphical Models (PGMs) have already been shown to be an adequate framework for managing uncertainty based on probability theory and are suitable for capturing dynamic relations among random variables. Under this scheme, brain regions can be thought of as random variables and the strength of the connectivity between them can be encoded in the links of the PGM. However, classical PGMs are based purely on conditional independence relations and although a special set of them attempt to make a causal interpretation, this subset still requires to have a large set of samples as well as

to satisfy a set of assumptions about the modelled world.

Through this document, a new methodology for learning causal PGMs capable of dealing with the constraints described will be developed and hypothesized to be computationally feasible. Collaterally, a proposal for the analysis of effective connectivity in the adult human brain from fNIRS will be described, considering the set of intrinsic characteristics of fNIRS data. The new neuroimage analytical tool is the byproduct of a yet unresolved computational problem; i.e., the establishment of a model for causal discovery considering only a reduced set of observations and not relying on strong assumptions demanding unrealistic overall control of the variables. These among others issues will be further discussed through this proposal.

1.1 Motivation

Understanding how variables are related in a system is one of the main aims in many fields across science. Often, researchers are interested in learning how these variables came to take their current values and what would happen if the system was disturbed. Senior researchers such as Clive Granger (Granger causality inventor and Nobel Prize 2013), Clark Glymour, Peter Spirtes and Richard Scheines (inventors of FCI algorithm) and Judea Pearl (inventor of Bayesian networks, the calculus framework for causal inference and Turing Award 2011), among others, have addressed their main research to the study and development of mathematical and graphical causal modelling.

Notwithstanding, despite the advances in the application of causal discovery methods in many natural phenomena, several problems such as a large set of variables, common confounders or latent variables and a small set of samples still remain under study in order to obtain safe practical tools for modelling these scenarios. PGMs are raising as promising tools for modelling of causal relations. However, their capacity of expressing causality is still ballasted by the demands of strong assumptions that often strictly limit their applicability to non-synthetic scenarios. These assumptions must be challenged by more aggressive computational strategies. For instance, the TETRAD project[1] led by Spirtes at the Carnegie Mellon university and supported by NASA is aimed at discovery causal statistical models. Current TETRAD research is centred on the extent to which limiting assumptions can be relaxed, thereby extending and investigating the extent to which the search procedures can be made more reliable on small samples. In (Spirtes, 2010), Spirtes established eleven remain open problems regarding the area of causal modelling. Two of them are: improve efficiency and efficacy of search algorithms and adding or relaxing simplifying assumptions. The first it is aimed to find a method to incorporate different kinds of background information while the founded structures can be more reliable at small sample size. The second one is related to identify assumptions that might be weaker in order to be met in other domains or to check for stronger

[1] `www.phil.cmu.edu/tetrad/`

assumptions that can allow for stronger causal inferences. Today, many of the research in causal discovery has been turned to investigate how algorithms for causal discovery can be introduced to more domains and practical applications by taking advantage of the knowledge domain (incorporating background information) and alleviating the causal requirements (assumptions) to match with natural scenarios.

Using causal PGMs is appealing but there are still some computational challenges such as the learning with a small set of observations, integrate theoretical disturbs and the analysis of the extent to which the required assumptions could be relaxed. The insufficient set of observations hinders the discovery of complete structures since some methods make use of statistical tests and a few other cannot deal with the potential selection bias yielded by small samples[2]. Moreover, according to the domain circumstances, it is necessary to relax those assumptions that demand the overall control of the variables in the system, as well as to determine a suitable way to integrate theoretically disturbs over variables.

1.2 Justification

The Brain Research through Advancing Innovative Neurotechnologies (BRAIN) created in 2013 represents the greatest project to date for brain understanding. With an initial funding of $100 million per year, it is expected that the commitment of National Institute of Health will be about $300 million per year in the next years (through Advancing Innovative Neurotechnologies , BRAIN). The BRAIN's initiative main goal is to "accelerate the development and application of new technologies that will enable researchers to produce dynamic pictures of the brain that show how individual brain cells and complex neural circuits interact at the speed of thought". Accordingly to this, members of the BRAIN Initiative working group consulting with scientific community determined seven areas of high priorities. The fourth area **demonstrating causality**, is described as "link brain activity to behaviour with precise interventional tools that change neural circuit diagrams". This expresses the need of the creation of modelling tools able to capture neural circuits which allow exploring the responses of interventions over such circuits.

This research is framed by this goal which is aimed to explore the use of interventions in a circuit and explore the caused responses. Similarly, in this proposal the use of interventions in the computational model will be explored in order to improve the circumstances of which a set of causal relations in a system can be discovered in the PGM. Equally important are the application of the learning of causal relations in other areas such as medicine, social sciences and public policies. In these cases by learning a causal model, researchers can make inferences as well as interrogate the system under study in order to answer questions that will be unable to answer by other analysis tools.

[2]Large samples can also be affected by selections bias, although naturally they are better prepared to escape it

1.3 Problem statement

This research will address the phenomenon of retrieving the causal structure E of a network M from a set of indirect observations and the exploitation of knowledge of the problem domain by means of probabilistic relations. More specifically, the network M relates a set of random n-dimensional variables $V = \{v_i\}$ through some directed arcs $E = \{e_k; e_k = (v_i, v_j) \text{ with } v_i, v_j \in V\}$, such that $I(M) \subseteq I(P)$ where $I(\cdot)$ is a set of independencies and P is some distribution drawn from a set of multivariate spatiotemporal observations of the form $o(h; x, t, \theta)$ of effect multivariate random variables where h represents the observed multivariate signal at $< x, t, \theta >$ with x represents the spatial location, t represents the temporal location and θ encodes the experimental unit. In particular, this research will explore the particular conditions where:

- The observations $O = \{o_1, \ldots, o_n\}$ are indirect non-linear consequences of the causal variables in V. Although strictly, any o_i might express information arising from a combination of the occurrence of any subset of V, in this research it will be assumed that there is one o_i uncontaminated for each v_i, i.e. a one-to-one relation (o_i, v_i), $N = I$ can be established with N and I being the cardinality of the sets O and V respectively.

- The set of vectorial observations O available has a low cardinality, whereby "low" is not any specific quantity but only suggests that they might be insufficient to adequately characterize the underlying joint probability distribution of the variables in the set V . To deal with the low cardinality of available observations, a mechanism to identify a suitable subset of V to perform interventions to the network will be proposed, in order to generate a set of additional interventional data $\widehat{O}$ which in combination with O is aimed to output a version of M in which the number of discovered causal relations shall tend to increase.

- The problem domain might provide preliminary clues about the network causal structure E. These can be expressed as mathematical constraints guiding the structure learning process. So according to the problem at hand, these can be incorporated as a priori information to disentangle undefined discovered relations.

1.4 Research questions

1. *Given a multivariate data sampled from a causal process, how can the structure of a PGM encode the causal relations involved in such process when considering structural constraints?*

2. *Under what conditions a causal PGM can be learnt with a greater PAG accuracy considering a reduced subset of samples than a PGM model learnt with a dataset*

in the large sample limit? And how does the model fitness relates to the dataset sample size?

3. *Having learnt a multivariate causal model under certain assumptions and with the available number of observations, how can the learnt structure be further improved by intervening the PGM? Specifically, how such interventions have to be made in order to elucidate the causal direction of remaining undefined directions?*

1.5 Hypothesis

As a result of the aforementioned research questions it is possible to establish the general hypothesis of this research:

Considering a small set of samples given by a set of observations of the system under study a learning algorithm for causal discovery incorporating the background knowledge of the problem domain as well as combining observational along with interventional data, yields a PGM structure with greater number of invariant marks in contrast with the number obtained if only observational data were considered, such that the true causal structure of the system is more faithfully captured (i.e. maintaining at least the goodness of fit).

1.6 Objectives

1.6.1 General objective

Develop and validate a methodology to learn a causal PGM which considers:

1. a limited number of observations,
2. the combination of observational and interventional data, and
3. accepting a prior domain knowledge.

1.6.2 Specific objectives

These objectives are oriented to give a set of solutions related to the research questions presented in Section 1.4.

1. Development of a structure learning algorithm for generating a PGM which captures the underlying causal relations among a set of variables. The appropriateness will be assessed by means of the accuracy of the solution generated, i.e. the partial ancestral graph (PAG)(See next section).

2. To develop a mechanism for incorporating background knowledge from the problem domain, in a form of a set of mathematical restrictions into the causal discovery algorithm. The effectiveness of this mechanisms will be evaluated in terms of the average of the number of directed acyclic graphs (DAGs) implied by the class of equivalence yielded by the causal discovery algorithm in contrast with the number obtained when background knowledge is not included.

3. To analyse the theoretical requirements for causal discovery in order to identify to which extent an assumption of the computational model can be relaxed. In this case, the evaluation it is expected to be an analytical form of the discussion.

4. To establish a mechanism capable of selecting a set of target variables over which performing interventions in order to increase the number of invariant features. This mechanism will be verified by comparing the *partial ancestral graph* (PAG) accuracy of the obtained model with the PAG accuracy in the model where interventions are randomly performed.

5. To define a strategy for the combination of observational and interventional data obtained by the intervention over a set of target variables. This stage will be evaluated as a part of the learning algorithm.

6. To verify and validate a solution based on PGMs for causal analysis with the restrictions stated in the main goal and applied in the domain of neuroscience. According to the designed fNIRS experiment, a well-known neural circuit shall be recovered and interpreted.

1.7 Validation

The main aim is to evaluate the extent of how well a learned causal PGM is capturing what is suppose to do. In order to do so, the learnt structure will be compared with a synthetic ground truth in terms of the **PAG accuracy** which involves the set of edges of the structure whether considering a complete edge or considering its end-points (marks). The basic idea is to generate several causal scenarios as ground-truth structures. Once the structures are generated, several observations will be sampled from them and will be used as input datasets. Next, a network will be constructed for each dataset and finally, the average PAG accuracy will be obtained from the individual PAG accuracy. The Figure 1.1 illustrates the general idea of the assessment.

The accuracy can be obtained by counting the number of edges either correct, erroneous or missed (Claassen & Heskes, 2012; Colombo *et al.*, 2012). More precisely, the accuracy of the recovered structures will be calculated in terms of the number of pairs of variables having the correct direction if they are connected, or its non-connection in another case. For the sake of clarity, two variables X and Y are correct connected if

$X \to Y \in G_{true}$ if and only if $X \to Y \in G_{output}$, and two variables are considered correct non-connected if $X \nrightarrow Y \in G_{true}$ if and only if $X \nrightarrow Y \in G_{output}$ (Eberhardt, 2007), where G_{true} and G_{output} are the real and the output structures respectively and $\nrightarrow$ means the absence of the link.

Given that the output of the learning algorithm will be an equivalence class (e.g. a PAG), consider whether an edge is correct or erroneous may be daring. Since an edge is defined by two marks, these edge mark can be either a tail ($-$), an arrowhead ($>$) or a circle ($\circ$). Therefore, just as the total correct edges will be obtained, the average of correct edge marks will be obtained as another instance representing the PAG accuracy.

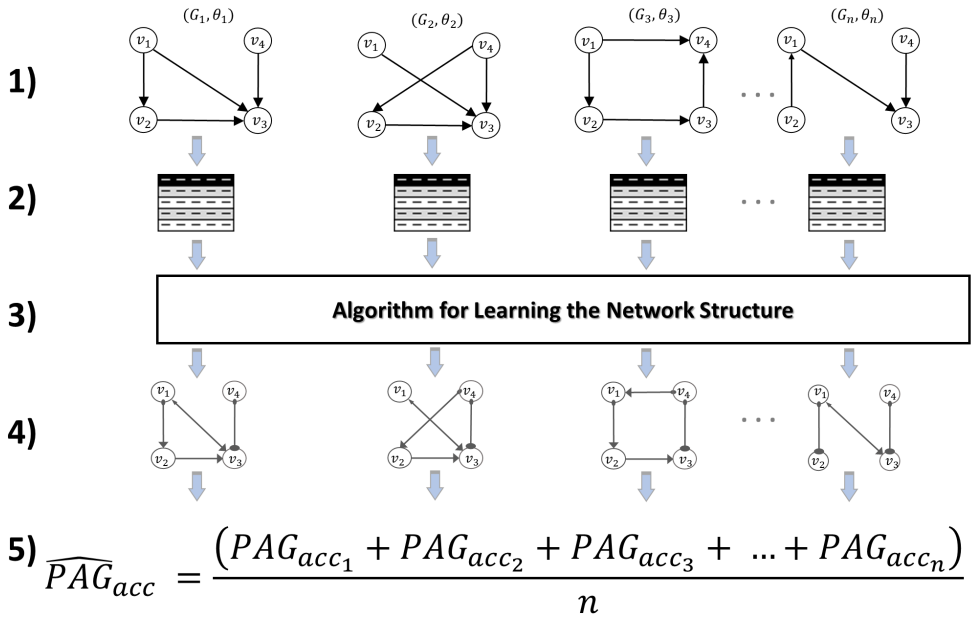


Figure 1.1: Diagram of the assessment of the algorithm. From top to bottom, 1) a set of predefined structures, 2) datasets obtained by sampling the predefined structures, 3) the learning algorithm, 4) respective learnt structures, and finally 5) the average PAG accuracy ($\widehat{PAG}_{acc}$) computed from the individual PAG accuracy (PAG_{acc_i}) for each structure.

These synthetic scenarios will be used as input datasets to FCI algorithm which is *de facto* the algorithm in the state-of-the-art for benchmarking when we deal with the discovery of a causal structure in PGMs. Finally, as a re-sampling method the K-fold validation technique will be used.

In terms of the application, considering a set of fNIRS neuroimaging from an experiment in neuroscience from a cohort of subjects performing a specific motor or cognitive task, the algorithm for learning causal PGMs will be aimed to recover the underlying active neural circuit involved in the task. Thus, the validation stage for this application will be the recovery of a well-known neural circuit and make a nomological validation

about the recovered circuit and the reported one in the literature. Such validation will be done since the true *in-vivo* neural circuit is not yet available, however, an inferred well-accepted circuit is available.

1.8 Contributions

Discovering the causal structure of a system of variables is not an easy task. Certainly, the existing methods have contributed to a specific domain of their respective areas. In this thesis, we address the **learning of the causal structure** between a set of variables from a probabilistic approach. In order to do that, an algorithm for the learning of a probabilistic model is proposed, which can be able to discover and represent a set of causal relations between variables. Furthermore, we will explore specific techniques in the PGM learning field which allow obtaining more informative structures, i.e., unravelling most of the undefined links.

In this sense, we expect to have three main contributions and one side contribution. The first three will impact to the learning field in computer science, specifically, in the learning of causal models as well as the strategies for combination of different data (experimental and observational). The remain contribution will impact in the neuroscience field by means of the recovery of the brain effective connectivity using fNIRS datasets.

Next, the expected contributions are listed:

- An algorithm for the learning of causal structures able to deal with datasets where the number of samples is insufficient to define the directionality of the relations.
- An algorithm able to select a set of target variables to perform experimental disturbances and combine it with observational data in order to elucidate the directionality of unresolved links.
- A mechanism to incorporate an efficient use of a priori information into a causal discovery algorithm.
- The recovery of effective brain connectivity from fNIRS data by means the proposed learning algorithm.

1.9 Publications

As part of the initial work developed and their results during the first year of the PhD., the following publication was submitted:

- **Samuel Antonio Montero-Hernández**, Felipe Orihuela-Espina, Javier Herrera-Vega and Luis Enrique Sucar, "Causal Probabilistic Graphical Models for Decoding Effective Connectivity in functional Near InfraRed Spectroscopy", submitted to the 29th International FLAIRS Conference, to be held in Florida, USA in May 2016.

Chapter 2

Background

2.1 Causality

Science, in general, is concerned with elucidating relations, whether predictive or explicative, among events or variables (Pearl, 2009b). Establishing causal (explicative) relations is one of the main aims in several disciplines across science. Notwithstanding, the term *causality* (often referred as *causation*[1]) entails many related but different tasks. Although these tasks are on the basis of a causal system, some of them are aimed to make inferences instead of to obtain the underlying causal model. Figure 2.1 illustrates the level of tasks in causation by means of hierarchical blocks. The bottom block represents a set of causal relations within a causal system over which different areas of analysis can take place, moreover instances of specific tasks are shown at the top block. This thesis is framed by the causal discovery area and specifically within the structure learning algorithms.

The definition of *causality* considered in this thesis is the one given by Spirtes *et al.*:

We understand causation to be a relation between particular events: something happens and causes something else to happen. Each cause is a particular event and each effect is a particular event. An event A can have more than one cause, none of which alone suffice to produce A. An event A can also be overdetermined: it can have more than one set of causes that suffice for A to occur. We assume that causation is (usually) transitive, irreflexive, and antisymmetric. That is, i) if A is a cause of B and B is a cause of C, then A is also a cause of C, ii) an event A cannot cause itself, and iii) if A is a cause of B then B is not a cause of A. (Spirtes *et al.*, 2000, Ch 3.2).

Once the area of study for this thesis has been framed it becomes necessary to detail the landscape in the causal discovery area. Causal relations can be identified in two ways: controlling *all* possible factors present in a given system, perturbing the potential

[1]We will use both terms interchangeably herein.

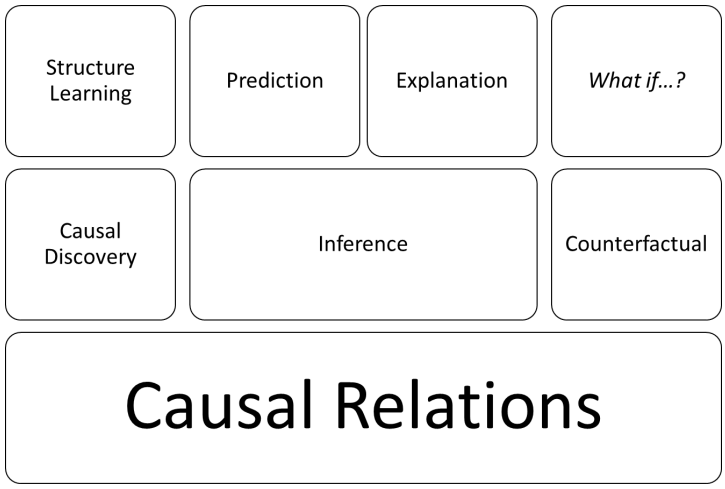


Figure 2.1: Levels of tasks in causality. On the basis of a set of causal relations within a causal system (bottom), three area of analysis in causation (middle) and specific task address by each area (top).

cause variable and measuring the effect obtained in the dependent variable or, by means of the analysis of the causal precedence of one variable over another.

Once a causal relation is discovered, it is possible to perform operations such as predict future behaviours (inference) or answer questions of the type *what if ...?* (counterfactuals). In some situations, only limited observations of the system are available i.e., reduced samples from a set of variables, and no additional information about the causal structure of the system is available. In that case, the researcher faces the task of recovering the causal structure of the system in disfavourable conditions. Recovering the causal structure of a system only from observational data is not possible (Pearl, 2000) since the need of controlling variables is fundamental. However, under certain circumstances although some variables cannot be controlled at least these can be measured, and thus, the structure of the model can be partially recovered (Scheines, 1997). Several aspects that hinder the retrieval of the causal structure of the system include, but not limited to, a large number of variables, small sample sizes, potential unobserved variables and bias selection when sampling.

According to the above, two major problems in causality can be stated: structure discovery and causal inference. Indeed, Judea Pearl (Pearl, 2000) establishes two principal questions:

1. *What empirical* ***evidence*** *is required for legitimate inference of cause–effect relationships?*

2. *Given that* ***we are willing to accept causal information about a phenomenon****, what inferences can we draw from such information, and how?*

The first question regards to the "what" and "how" to use the available data in order to capture causal relations. The second question highlights that a causal model is assumed to be an adequate estimate of the true phenomenon, so that may be able to afford some inferences.

One approach to address causal relations is the so-called *probabilistic causality*. This approach obtains causal information and makes causal inferences on the basis of probability tools. This may sound somehow limited because we know that probability deals with some kind of uncertainty and we would like to think that causal relations are deterministic relations by nature. However, when we observe daily phenomena, in most of them we are only able to determine a series of causes and effects to some extent. Sometimes, considering only the information at hand, we cannot state that event A is causing an event B and also ensure that the absence of A implies the not occurrence of B. Hence, it is most common to state that knowing information about A makes more likely that B occurs. In some sciences, there is more interest to know the extent or strength of a causal relation instead of its existence or absence (Pearl, 2009b).

In this work, the problem of estimating the causal structure of a phenomenon of interest, related to the first of the two major problems in causality, will be addressed.

2.2 Probabilistic graphical models

A probabilistic graphical model is a compact representation of a joint probability distribution. It provides a framework for managing uncertainty based on probability theory in a computationally efficient manner (Sucar, 2015). A natural way to represent the dependence and independence relations between a set of variables is using graphs, where dependent variables are connected, and the independence relations are implicit in this graph (not connected).

A probabilistic graphical model is specified by two aspects: a graph, $\mathcal{G}$ defining the structure of the model, and a set of local functions that define the parameters θ. The joint probability implied by the PGM structure is obtained by the product of the local functions:

$$P(V_1, V_2, \ldots, V_n) = \prod_{i=1}^{n} \theta_{V_i}$$

PGMs can be classified according to three dimensions: directed or undirected, static or dynamic and probabilistic or decisional. There are two major tasks for a PGM: learning and inference. Learning task consists in estimating the structure and parameters of the model given a dataset. On the other hand, inference deals with answering probabilistic queries by obtaining the conditional or marginal probability distribution of a subset of variables.

The research proposed here focuses (mainly) on the learning task, specifically for directed (causality involves direction) and probabilistic model (only random variables

will be included rather than decision or utility variables).

2.2.1 Graphs

There are some mathematical structures which can represent causal relations in an intuitive and explicit way, one of these are graphs. Graphs may be directed or undirected, may contain cycles or even may be mixed (a combination of directed and undirected). They have two basic set elements: nodes and edges. Formally:

Definition 2.2.1. A **graph** $\mathcal{G}$ is a pair (V, E) where V is the set of nodes $\{V_1, \ldots, V_n\}$ and E is the set of edges consisting of pairs of the form (V_i, V_j). If an edge is directed it is represented as $V_i \to V_j$, and if it is not directed then it is represented as $V_i - V_j$, as well if the direction is not relevant then it is represented as $V_i \rightleftharpoons V_j$.

If all edges in a graph $\mathcal{G}$ are directed then $\mathcal{G}$ is a **directed graph**. A graph may contain *paths*, *trails*, and *cycles*. A path is a sequence of nodes of the form $V_i, V_j, \ldots, V_n$ each of one connected by an edge with its predecessor. A trail is a path of the form $V_i \rightleftharpoons V_j \rightleftharpoons \ldots \rightleftharpoons V_n$ and a cycle is a directed path of the form $V_i, V_j, \ldots, V_n$ where $V_i = V_n$. A special type of graph is the **directed acyclic graph** (DAG).

Definition 2.2.2. A DAG $\mathcal{G}$ is a pair (V, E) where V is the set of nodes $\{V_1, \ldots, V_n\}$ and E is the set of edges consisting of *ordered*[2] pairs of the form (V_i, V_j) which implies $V_i \to V_j$ (a connection *from* V_i *to* V_j) where $i \neq j$ and also have no cycles.

A DAG is of special interest in the study of causal relations due to its properties: directed edges and no cycles. Directed edges allow establishing the flow of information between variables, whereas the absence of cycles avoid having a variable as cause and effect at the same time[3].

Given a DAG $\mathcal{G} = (V, E)$, whenever $V_i \to V_j \in E$, it is said that V_j is the child of V_i and that V_i is the parent of V_j in $\mathcal{G}$. Additionally, when $V_i - V_j \in E$, it is said that V_i is a neighbor of V_j (and vice-versa). It is said that V_i and V_j are adjacent whenever $X \rightleftharpoons Y \in E$. The notation $pa(V_i)$ denotes the parents of V_i, $ch(V_i)$ to denote its children, and $nb(V_i)$ to denote its neighbors. In a causal sense, the link $V_i \to V_j$ implies that V_i is causing to V_j, and in a temporal sense implies that V_i precedes V_j. The needed implications for different interpretations are discussed later.

2.2.2 Probability

Using the concept of probability we understand a certain extent of belief assigned to a specific event. Assuming Ω as the set of all possible outcomes of all events defined on a space and defining two events A and B as subsets of Ω, a basic expression in probability

[2]Mathematically, a particular sorting of a given set of objects, in this case the pair (V_i, V_j).

[3]A property that is desired only when we are not interested in mutual causation or feedback processes.

theory is the concept of *conditional probability* $P(A|B)$ which represents the belief of A given that B has occurred. In this sense, for each level or subsets of outcomes in B a probability distribution of A is defined. If $P(A|B) = P(A)$ then A and B are **independent** because knowing about B does not affect the belief of A.

If $P(B) \neq 0$ the conditional probability $P(A|B)$ is defined as:

$$P(A|B) = \frac{P(A,B)}{P(B)} \tag{2.1}$$

From Eq. 2.1 it follows

$$P(A,B) = P(A|B)P(B) \tag{2.2}$$

Generalizing for k events, if $A_1, A_2, \ldots, A_k$ are events, such that $P(A_1, A_2, \ldots, A_k)$ represents the **joint probability distribution** over variables $A_1, \ldots, A_k$ then the *chain rule* is defined as

$$P(A_1, A_2, \ldots, A_k) = P(A_1)P(A_2|A_1)\cdots P(A_k|A_1, A_2, \ldots, A_{k-1}) \tag{2.3}$$

In other words, it is possible to obtain the probability of a series of joint events, as a function of the probability of the first event, the probability of the second event given the first, and so on, regardless the sequence order.

Having the concepts of conditional probability and joint probability, next the *conditional independence* is defined:

Definition 2.2.3. Let $V = \{V_1, V_2, \ldots, V_n\}$ be a finite set of variables. Let $P(\cdot)$ be a joint probability distribution over the variables in V and let X, Y and Z be three subsets of variables in V. The sets X and Y are said to be independent given the set Z if

$$P(X|Y,Z) = P(X|Z) \text{ whenever } P(Y,Z) > 0 \tag{2.4}$$

In other words, learning about Y does not provide further information about X when Z is known. The next notation is introduced:

$$(X \perp\!\!\!\perp Y|Z) \text{ iff } P(X|Y,Z) = P(X|Z) \tag{2.5}$$

where $\perp\!\!\!\perp$ stands for *independent*. The next section introduces one type of PGM for which some structure discovery algorithms have been developed.

2.2.3 Bayesian networks

Bayesian networks (BN) are structures whereby a joint probability distribution can be coded. BN are compact representations of a joint probability distribution over a set of variables. For example, consider a joint probability distributions over a set of variables $(V_1, V_2, \ldots, V_n)$ and suppose that each one has two values, then a table of $2^n - 1$ parameters is needed to represent the joint distribution. In contrast, a BN reduces the

total number of parameters capitalizing on the in/dependence relations described by the set of links. Such relations allow decomposition of large global distributions into small local distributions of less variables (Koller & Nir, 2009). Consequently, a substantial reduction of memory space is achieved.

Formally, a Bayesian network $\mathcal{B}$ consists of a pair $(\mathcal{G}, \theta)$, where $\mathcal{G}$ is a DAG and θ stands for the parameters needed for encoding a set of local probability distributions. Besides, the pair (V, E) conforming $\mathcal{G}$ has a special meaning. The set V contains all the variables presents in the joint probability distribution and each edge $(V_i, V_j) \in E$ means that V_i and V_j are not independent. The joint probability distribution coded by a BN is calculated as

$$P(V_1, V_2, \ldots, V_n) = \prod_{i=1}^{n} P(V_i | pa(V_i)) \tag{2.6}$$

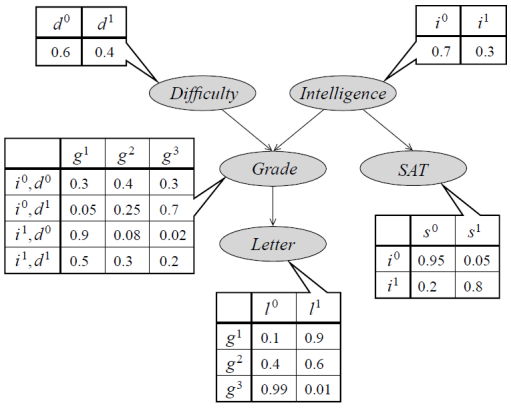


Figure 2.2: An example of a BN extracted from (Koller & Nir, 2009). Student network with the conditional probability distribution associated with each variable. Superscripts 0 and 1 means low and high values in variables *Difficulty, Intelligence, SAT* and *Letter* and g^1, g^2, g^3 stands for "A", "B" and "C" score in variable *Grade*.

An example of a BN is shown in Fig 2.2. This Student network of the example is a complete BN from which it is possible to make some probabilistic queries, such as

- How is affected the probability of *SAT* once we know that *Intelligence* is *low* (i^0)?
- What is the likelihood of high intelligence in a student given that he obtained a C score in the final grade?

These type of questions are referred to as causal reasoning and evidential reasoning respectively. However, it is necessary to have the BN structure and its parameters in order to answer causal and evidential questions. The BN structure can be obtained in three ways: (1) designed by an expert, (2) guessing-and-testing or (3) by means of

learning structure algorithms. The first two ways are unfeasible when an expert is not available or when there are several variables resulting in a large number of possible networks[4]. The algorithms for learning the structure of the network can be coarsely divided into two approaches; constraint-based and score-based methods.

Constrain-based algorithms identify dependence relations between variables by means of a statistical test conditioned on a subset of variables. The basic idea is starting with a completely connected graph, testing whether each variable and its neighbours are independent or not. For instance, under the null hypothesis that the $X, Y \in V$ set of variables are independent, a statistical test evaluates if X and Y are conditionally independent or not. If the result of the test is statistically significant then the null hypothesis is rejected and thus, X and Y are regarded as not independent, which implies removing the link between X and Y in $\mathcal{G}$. Different statistics can be used based on the type of data. With the assumption of Gaussian distribution Fisher z-transformation can be applied (Kalisch & Bühlmann, 2007), in case of discrete data the G-test can be used (Neapolitan, 2003), for categorical variables Pearson χ^2 is often used (Koller & Nir, 2009) and the G^2-test can be used for binary variables (Neapolitan, 2003).

In score-based methods, it is necessary to explore the space of possible structure graphs by searching for the structure that better fits the evidence. This process is computationally expensive because the number of possible graphs grows quickly as the number of variables increases. The number of total DAG $f(n)$ given n variables is defined in a recursive form (Robinson, 1973, 1977) in Eq. 2.7.

$$f(n) = \sum_{i=1}^{n} (-1)^{i-1} \binom{n}{i} 2^{i(n-i)} f(n-i) \text{ with } f(0) = 1 \tag{2.7}$$

Using Eq. 2.7, the number of DAGs for $n = 4, 5, 6$ are $543, 29281$ and 3781503 respectively, thus, an exhaustive search quickly becomes computationally unfeasible. Consequently, several heuristics have been proposed in order to search for a local optimum structure over the search space. These methods attempt to obtain a reasonable trade-off between the exploration and exploitation of the space of structures. Turning the problem of finding the sub-optimal structure that best fits the data into an optimization problem, score-based methods evaluate the fitness of the structure by means of measures such as the maximum likelihood (ML), the Bayesian information criterion (BIC), the Bayesian score (BD), and the minimum description length (MDL) criterion (Sucar, 2015).

Some algorithms limit the search process to the task of evaluating the equivalence class of a partition of the structure space instead of all the structures contained in that partition (Chickering, 2002). Before giving the definition of an equivalence class, an equivalence relation is defined.

Definition 2.2.4. An equivalence relation on a set X is a subset of $X \times X$ where $x \sim y$ means that (x, y) are related, and $\sim$ is an equivalence relation if and only if, is reflexive

[4]In this case a strategy of search is needed.

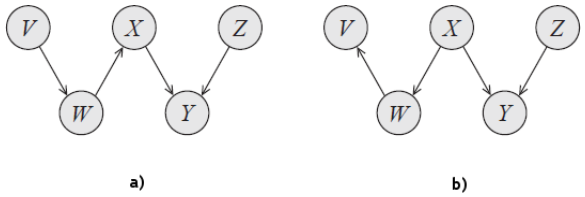


Figure 2.3: Equivalent Bayesian networks, both a and b encode the same conditional independence relations.

($x \sim x$), symmetric ($x \sim y$ implies $y \sim x$) and transitive (if $x \sim y$ and $y \sim z$ then $x \sim z$), with $x, y \in X$.

In other words, Definition 2.2.4 states that two elements of a set are equivalent related if both are in the same subset, such that all the subsets of the originally set are disjoint (their intersection is empty) and the union of all subsets conforms the original set. Then, it is said that a subset of this form is formally called an *equivalence class*.

Definition 2.2.5. An **equivalence class** is defined as a subset of the form $[a] = \{x \in X : x \sim a\}$ where a is an element of X and the notation $x \sim y$ means that there is an equivalence relation between entities x and y.

An equivalence class as defined in the Definition 2.2.5, implicitly states that such equivalence class forms a partition over the set that is defined. Also, note that only it is required an equivalence relation, thus for two elements $x, y \in X$ to be equivalent depends on how the set X was partitioned. Classical BNs define the equivalence relation over the set of conditional independence. For instance, in Fig. 2.3 BNs a and b are equivalent according to the set of conditional independence relations encoded by each one. In order to understand the implication of this equivalence, it is necessary to define the concept of *Markov Compatibility*.

Definition 2.2.6. If a probability function P admits the factorization of Eq. 2.6 relative to a DAG $\mathcal{G}$, it is said that $\mathcal{G}$ represents P, and also that $\mathcal{G}$ and P are compatible, or that P is **Markov compatible** to $\mathcal{G}$.

Given that two equivalent BNs encode the same independence relations underlying in a probability distribution P, it is convenient to have a way to list all the conditional independence relations in a DAG. In 1988, J. Pearl proposed a convenient way to read off the set of conditional independence in a DAG $\mathcal{G}$ known as the *d-Separation*[5] (Pearl, 1988).

Definition 2.2.7. A path p is said to be *d-separated* (or blocked) by a set of nodes Z, iff

[5]The d denotes *directional*.

1. p contains a chain $i \to m \to j$ or a fork $i \leftarrow m \to j$ where the middle node m is in Z, or

2. p contains an inverted fork (or collider) $i \to m \leftarrow j$ such that the middle node m is not in Z and such that no descendant of m is in Z.

A set Z is said to *d-separate* X from Y iff Z blocks every path from a node in X to a node in Y. The Table 2.1 illustrates the meaning of Def. 2.2.7.

Path	$W \subseteq Z$	$W \subseteq Z$
$X \to W \to Y$	✓	✘
$X \leftarrow W \leftarrow Y$	✓	✘
$X \leftarrow W \to Y$	✓	✘
$X \to W \leftarrow Y$	✘	✓

Table 2.1: Examples of *d-Separation* ✓ indicates the flow of information and ✘ implies not flow of information. When set W is unobserved ($W \nsubseteq Z$) information flows in the first three paths, whereas when W is observed ($W \subseteq Z$) information solely flows in the fourth path.

Thus, two DAGs can be evaluated to be equivalent by comparing the set of conditional independencies obtained by means of the *d-Separation*, resulting in an associative interpretation between variables rather than a causal interpretation.

2.2.4 Causal PGMs

The interpretation of PGMs such as classical Bayesian networks in the independence sense does not necessarily means causality (Pearl, 2009b). However, it is possible to give a causal interpretation to DAGs. To afford such interpretation, the meaning of the relation implied by the link $X \to Y$ must be extended. From now, the relation $X \to Y$ means a **causal** relation, such that X is a *cause* of Y. This leap in semantics demands further explanation and mathematical support. Causal models need to be able to represent *sudden changes* in the system. Sudden changes express disturbs in the variables behaviour by external factors regardless its prior probability distributions. The representation of such changes by means of a suitable semantic allows to answer different causal questions. The most common operation to represent these changes is the *intervention*. An intervention on any specific variable leaves a causal effect as a result. One elemental question is how to measure the causal effect of one variable over another, or over the complete system. Basically, a causal effect is the response of one element of the system given the intervention on its direct causes. Another type of question once the structure of a causal model is known, is the *counterfactual* question, in which it is

possible to infer about the behaviour of a variable and its consequences if it would have had a different value instead of the observed one.

The simplest form of intervention is to force an element of the system to take some specific value (for instance force X_i to take the x_i value), then compute the new probability distribution from the modified system. Such an intervention implies isolating X_i from the influence of other variables so that the variable X_i is only affected by the new forcing mechanism. Formally, this intervention will be denoted as $do(X_i = x_i)$, replacing the classical probability measure $P(X_i = x_i)$ such that a new system is obtained (Pearl, 2000). The behavior under the intervention is represented by the new system and when the probability distribution over another variable (say X_j) is computed, the causal effect of X_i on X_j denoted as $P(X_j|do(X_i = x_i))$[6] can be obtained.

Definition 2.2.8. Given two disjoint sets of variables, X and Y, the **causal effect** of X on Y denoted as $P(y|do(x))$ is a function from X to the space of probability distribution of Y. For each intervention of X, $P(y|do(x))$ yields the probability of $Y = y$ induced by removing from the model all the other influences to the variables in X.

A Bayesian network with a causal interpretation is called **causal Bayesian Network** (CBN for short). Formally, a CBN $\mathcal{C}$ consists of a $(\mathcal{G}, \theta)$ pair, where $\mathcal{G}$ is a DAG and θ stands for the parameters (as in classical BN) and each edge $(V_i, V_j) \in E$ means that V_i is a cause for V_j. Besides, it is necessary that $\mathcal{C}$ is compatible with all the probability distributions over V resulting from interventions on $Y \subseteq V$ where V is the complete set of variables and y is a constant value, such that:

1. The resulting probability distribution $P_Y(V_i)$[7] derived from intervention is Markov compatible with $\mathcal{G}$ DAG (see Def. 2.2.6).

2. The joint probability distribution derived from an intervention over the variables Y is calculated as $P_y(V) = \prod_{\{V_i \notin Y\}} P(V_i|pa(V_i))$.

3. The probability of all the variables that are part of an intervention is equal to one for which the value is set to if $V_i = y$ is consistent with $Y = y$ then $P(V_i) = 1$.

4. The probability of remaining variables $V \backslash Y$ is equal to the probability of each variable given its parents and must be consistent with the intervention, i.e., $P_y(V_i|pa(V_i)) = P(V_i|pa(V_i)), \forall V_i \notin Y$.

Note the unfortunate ambiguous notation common in literature, which does not allow unambiguous distinction of a Bayesian network $\mathcal{B} = (\mathcal{G}, \theta)$ from Causal Bayesian Network $\mathcal{C} = (\mathcal{G}, \theta)$ unless explicitly stated.

[6] For short, sometimes $do(X_i = x_i)$ is expressed as $do(x_i)$.

[7] $P_Y(V_i)$ stands for the probability distribution over V_i when performing interventions on Y.

2.2.4.1 Learning a causal PGM

In all previous situations, the existence of the causal model was assumed. However, this is not always the case and sometimes there are only data from a set of observations of the system. Hence, the task of learning a causal model from data emerges. Learning a causal model from data can be hard because:

- the dataset can be originated purely from observations [8],
- maybe there exists hidden or unobserved variables within the system, and/or
- depending on the process of sampling, there may only be a limited set of samples.

Although all previous issues hinder the discovery of the true causal structure, it is worth to mention how the size of the samples affects to the discovery process. First and foremost, the constraint-based algorithms for structure learning rely on statistical test, i.e., they need to deal with statistical estimators of the sample[9]. Recall that, PGMs encode the joint probability distribution of the sample data which is close to the real probability distribution of the system from which the sample was obtained (real P.D.). So, the joint P.D. represented by the PGM structure is expected to fit the real P.D. Such goodness-of-fit have been measured using the Kullback-Leibler distance (Kullback & Leibler, 1951) as in (Abbeel *et al.*, 2006) or by assessing the performance of the learnt model as in (Greiner *et al.*, 2013). In spite of how the goodness-of-fit is assessed, the authors coincide that having a great number of samples is more suitable to discover the true structure of a network. By *great number*, we do not refer to a specific quantity but to obtain a high value of goodness-of-fit. Yet some works were aimed to identify the bounds of the sample size (Zuk *et al.*, 2012), the truth is that a dataset with a limited set of sample has a negative impact on the structures that classical learning algorithms learn with respect to the confidence interval.

Most of the proposed algorithms to achieve learning of a causal model attempt not to be affected by these issues. In consequence, they rely on a set of assumptions (Druzdzel, 2009; Sucar, 2015; Spirtes *et al.*, 2000) being some of the most common:

1. **Causal Markov Condition**, whereby a variable is independent of its non-descendants given its direct causes (parents in the graph).
2. **Causal Faithfulness**, whereby there are no additional independencies between the variables in the model that are not implied by the causal Markov condition.
3. **Causal sufficiency**, whereby there are no common confounders of the observed variables in the model.

[8] Recall that it is simply impossible to guarantee causality from observational data alone (Pearl, 2009b)

[9] By *sample* we refer to the set of observations obtained from the causal system.

4. **Closed-world**, whereby the true model accounts for all relevant variables.

5. **Acyclicity of the true causal structure**, whereby there are no reflexive relations for all variables.

Sometimes, even with these assumptions current causal discovery algorithms are unable to find "the" accurate representation of the causal model. Instead, they provide an equivalence representation of the most likely models which share a set of invariant features. The invariant features relate elements such as the same skeleton, v-structures and edge marks (tails or arrowheads) of a set of DAGs. Such invariant features are represented via Ancestral Graphs (AGs) (Richardson & Spirtes, 2002).

Definition 2.2.9. AGs are an extension of DAGs in which bidirect arcs of the type $X \leftrightarrow Y$ and undirected arcs $X - Y$ are allowed.

Two properties rule the flow of information in an AG, which are:

1. If V_i and V_j are joined by an edge with an arrowhead at V_i, then there is no direct path from i to j.

2. There are no arrowheads present at a vertex which is an endpoint of an undirected edge.

Using an equivalence class of a set of DAGs or AGs, causal discovery algorithms represent their output model set. In some situations, algorithms cannot estimate all the causal relations accurately, because they find equivalent models that can explain the data equally well. The equivalence class for DAGs or AGs is represented by a *Partial Ancestral Graph* (PAG). In a PAG invariant marks are preserved, tails ($-$) and arrowheads ($>$) that appear in all members of the class are preserved and the non-invariant edge marks are represented by a circle ($\circ$). Formally (Claassen, 2013):

Definition 2.2.10. A **partial ancestral graph (PAG)** $\mathcal{P}$ is a mixed graph that represents the equivalences class $[\mathcal{M}]$ of an MAG $\mathcal{M}$, such that:

- it has the same skeleton as $\mathcal{M}$,

- all non-circle edge marks (arrowheads and tails) in $\mathcal{P}$ correspond to invariant marks in $[\mathcal{M}]$.

An example is illustrated in Fig. 2.4.

To summarize, the two principal tasks in the context of causality are *causal inference* and *causal discovery procedures*. The former is related to answering different kind of questions including predictions, interventions, and counterfactuals, whereas the later is concerned with methods for discovering and representing causal models (e.g., causal Bayesian Networks) from data. In short, it can be said that one of the main objectives in causal discovery is finding as many invariant features of the equivalence class as possible.

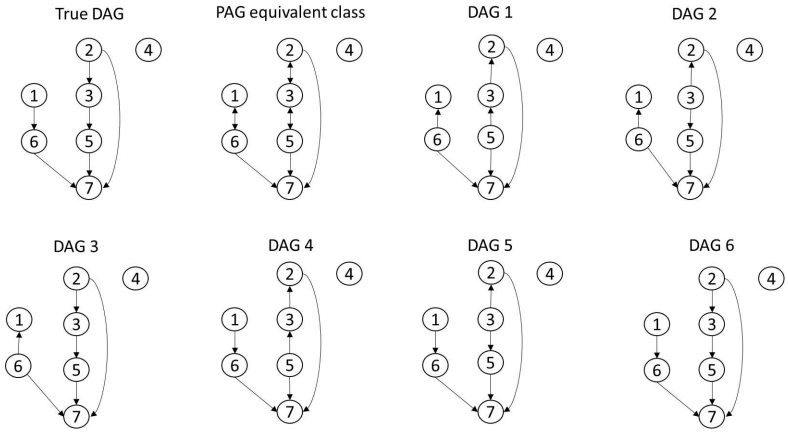


Figure 2.4: Example of equivalence class for a DAG: true DAG, PAG and the DAGs in the equivalence class Figure adapted from (Kalisch *et al.*, 2012).

2.3 Domain of application

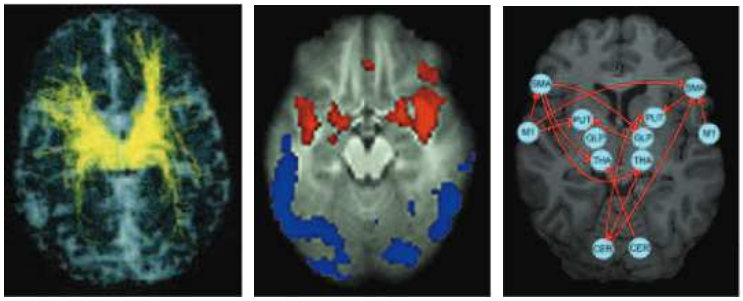
2.3.1 Brain connectivity

The human brain is the organ of the central nervous system, responsible for high cognitive functions such as memory function and controlling volitional motor skills (Frackowiak *et al.*, 2004). Many tasks performed by the human brain involves the reception of information (afferent sensory flow), the process (interneural flow) and the response to the information (efferent motor flow) perceived. When some process are performed, a series of connections take place in the brain.

There are three modalities of brain connectivity for the analysis of the different types of interaction: structural, functional and effective. The structural connectivity determines the physical links between neuronal elements. Functional connectivity identifies the statistical patterns between brain regions. And effective connectivity aims to capture the causal relations from which a brain region influences the activation or inhibition over another.

Tractography traced from data acquired from diffusion tensor imaging (DTI), is used to visualize the structural (anatomical) network. The full network of structural connections is known as the *connectome* (Sporns *et al.*, 2005), and decoding it, is currently one of the major challenges of science (Rosen *et al.*, 2010).

Functional and effective connectivity are explored with functional neuroimaging modalities such as functional magnetic resonance imaging (fMRI) (Li *et al.*, 2012; Rajapakse & Zhou, 2007), electroencephalography (EEG) (Achermann & Borbély, 1998), magnetoencephalography (MEG) (Kiebel *et al.*, 2009), positron emission tomography (PET) (Friston *et al.*, 1996) and functional near infrared spectroscopy (fNIRS) (Strangman *et al.*, 2002; Villringer & Chance, 1997) among others (Martin, 2014). The main difference between functional and effective connectivity is related to the nature of



(a) Structural connectivity (b) Functional connectivity (c) Effective connectivity

Figure 2.5: Examples of brain connectivity Figure reproduced from (Li *et al.*, 2012).

relationship sought. In functional connectivity associative relations are sought, whereas effective connectivity attempts to recover causal relations. Examples of the anatomical, functional and effective connectivity are illustrated in Fig 2.5.

Henceforth, we focus in the human brain at macro-scale, specifically in the design of tools for the analysis of effective connectivity.

2.3.2 Functional Near-Infrared Spectroscopy (fNIRS)

Diffuse optical measurements are commonly performed through three techniques: continuous wave, time-resolved and frequency-resolved. Continuous wave modality has good sampling rate although the penetration depth is minor in contrast with time-resolved and frequency-resolved techniques, however, is the lowest cost technique. Time-resolved and frequency-resolved have an accurate separation of the absorption and scattering of the emitted light, although they have some disadvantages such as poor penetration depth for frequency-resolved, and for time-resolved issues as high cost, stabilization, instrumentation size and sampling rate hinder their use [10].

A special case of DOI is continuous wave functional near-infrared spectroscopy (fNIRS). fNIRS measures the optical density changes encoded in the backscattered radiation resulting from near-infrared light previously irradiated onto the scalp. Next, using the modified Beer-Lambert law (Cope *et al.*, 1988) the changes in oxygenated (HbO_2) and deoxygenated haemoglobin (HHb) can be estimated. Both HbO_2 and HHb are the two dominant changing chromophores in the adult head, and are relevant indicators of the neural activity by means of the neurovascular coupling (Strangman *et al.*, 2002). fNIRS presents a set of advantages including portability, mildly robustness for motion artifacts, a good trade-off between temporal and spatial resolution and, importantly it is a non-ionizing and non-invasive method, permitting multiple scan sessions and feasibility

[10]For brevity the physical principles of diffuse optical imaging are not presented here, and the reader is directed to (Boas *et al.*, 2011) for a more detail treatment.

of continuous monitoring for extended periods.

In order to generate an optical neuroimage, the processes of *image formation* and *reconstruction* take place. The image formation refers to the physical process whereby the light interacts with the matter so that the radiation exiting the tissue can be sensed by a photosensitive device. The Fig. 2.6 shows a schematic of the image formation process. The image reconstruction process attempts to recover the optical parameters affected by the interaction between irradiated light and the biological matter.

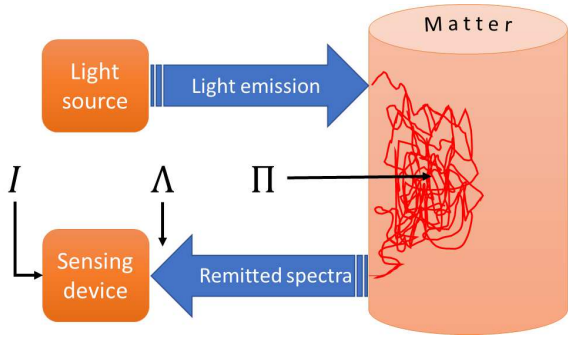


Figure 2.6: Schematic depiction of the image formation process in optical neuroimaging, where Π, Λ and I are the parameters, remitted spectra and image spaces respectively.

It is possible to define both processes mathematically as the mapping from the parameter space Π to the remitted spectra space Λ and then from there to the image space I, i.e., $G : \Pi \to \Lambda \quad F : \Lambda \to I \quad H : F \circ G$ with $\Pi \in \mathrm{R}^n \quad I \in \mathrm{R}^m$. Therefore, an element of an observation in the image space is defined as $i = F(G(\pi))$ with $\pi \in \Pi$. The process of image reconstruction is defined as the inverse process so that it is defined as $H^{-1} : I \to \Lambda$.

In optical imaging, the image of interest is the set of parameters present in the object. The inverse problem of the image reconstruction has a set of inherent issues hindering the recovery of the parameter space. The mapping G from parameter space to spectra space is not a monotonic function which renders the separation between chromophores an ill-posed problem. The so-called *metamerism* originated by the projection from the parameter space to a lower dimension (image space) causes to obtain an identical remitted spectra from different parameter configurations. Besides, the use of measuring devices involve the loose of information due to the suboptimal quantum efficiency, as well as the inherent noise introduced by these devices.

Those issues, as well as the characteristics of the fNIRS neuroimage data, have to be considered either to be included in the design of the model or analyse the set of assumptions that are true in these circumstances when modelling the effective connectivity as the topological and statistical properties of the dataset are altered.

Chapter 3

Related Work

This chapter is aimed at presenting the relevant works in the establishment and learning of causal relations. These are grouped into three blocks: causal models from a generic point of view, causal discovery algorithms and mechanisms for the combination of observational along with interventional data.

3.1 Causal models

A definition of a causal model was provided by Pearl (Pearl, 2009b, Def. 7.1.1).

Definition 3.1.1. A **causal model** is a triplet

$$M = \langle U, V, F \rangle$$

where U is a set of background (exogenous) variables that are determined by factors (variables) outside of the model, V is a set $\{V_1, V_2, \ldots V_n\}$ of variables, called endogenous, that are determined by variables $U \cup V$ in the model, and F is a set of functions $\{f_1, f_2, \ldots, f_n\}$ such that each f_i is a mapping from $U \cup (V \backslash U)$[1] to V_i and such that the entire set F forms a mapping[2] from U to V. In other words, each f_i tells the value of V_i given the values of all other variables in $U \cup V$, and the entire set F has a unique solution. Symbolically, the set of equations F can be represented by

$$V_i = f_i(pa(V_i), U_i),\ i = 1, \ldots, n\,,$$

Likewise, $U_i \subseteq U$ stands for the unique minimal set of variables in U sufficient for representing f_i.

A causal model M can be associated with a DAG $\mathcal{G}$ in which each node corresponds to a variable and the directed edges point from members of $pa(V_i)$ and U_i toward V_i. So

[1] "\" stands for the set difference operation, i.e., $A \backslash B = \{x : x \in A \text{ and } x \notin B\}$

[2] A mapping is a function that preserves the algebraic structure.

the graph $\mathcal{G}$ represents the influences of exogenous and endogenous variables over the other variables in V_i. In a nutshell, an instance of a causal model may be represented by a set of nodes depicting a set of variables (some of which take their values from factors outside of the system and the rest from factors within the system) and a set of links implying direct influences between variables.

Structural equation modelling (SEM) identifies causal relations between variables within a system by expressing such relations in terms of differential equations(Hoyle, 2012). SEM allows modelling theoretical constructs by including latent variables. Usually, structural equation models are represented by a path diagram (Wright, 1921) (see Fig. 3.1) in which background (unmeasured) variables are drawn as circles, endogenous (observed) variables are drawn as rectangles. Furthermore, arrows among variables represent the causal relations where factors at the tail are causing factors in the arrowhead. The strength of relations between variables is represented by regression or path coefficients in terms of weights. Bidirectional arrows represent correlational relations rather than causal relations.

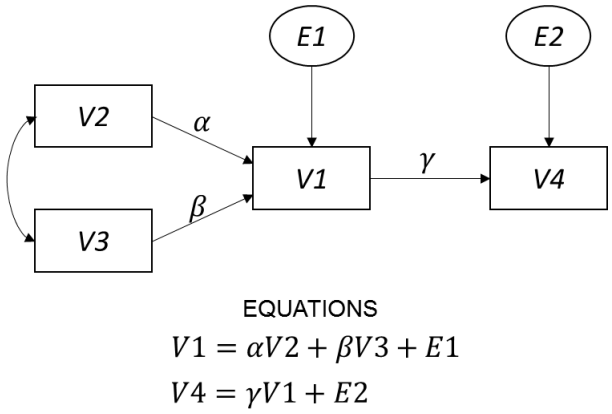


Figure 3.1: Example of an structural equation model. Figure adapted from (Schumacker & Lomax, 2004, p. 283).

Often, precedence between variables can furnish knowledge about causal relations. Based on the temporal precedence, Granger established a technique to determine whether an early variable is causing to other (Granger, 1969). In the temporal context, variables represent time series. According to Granger causality a variable with temporal dynamics e.g. a time series X_1 Granger-causes (or G-causes) another temporal variable X_2 if past values of X_1 help to predict X_2 in a better way than only consider the past values of X_2 alone. Granger's causality (G-causality) is formulated by linear regression of time series. The original form of G-causality presents some limitations as the required linearity and stationary of the signals, as well as being unable to deal with latent factors (i.e. does not account for the context).

Causal Bayesian networks (CBNs) briefly presented in the previous chapter are motivated from the limited interpretation of classical Bayesian networks regarding express

causal relations that are instead merely associational (Pearl, 2000). A key point in CBN is the set of assumptions made in order to deal with some aspects when causal relations are inferred from a set of observations. Moreover, the output of the algorithms for learning structure is a set of possible causal structures represented by an equivalence class. According to the set of assumptions on which relies, CBNs may be able to allow latent (unmeasured) and/or selection variables.

As shown above, there exist different approaches for recovering causal relations. Main differences among these methods are the representations of causal relations (path diagrams, differential equations, linear equations, graphs, autoregressive models), the type of data analysis (bivariate, multivariate) and whether they are confirmatory or exploratory (hypothesis-driven, data-driven). Importantly, every approach raises in a certain field of the science, so it can be stated that they attempt to capture the same construct although they tell the story dissimilarly. Mainly, the difference might occur due to the assumptions that are true in the area that they emerge, hence the importance of scrutinizing the characteristics of the problem domain.

3.2 Causal discovery algorithms

Within the PGM framework, there are several proposed algorithms for the learning of causal relations from data. Based on independence relations, these algorithms rely on a set of assumptions (commented in Section 2.2.4.1) whereby the recovered structure can be causally interpreted. The main algorithms and their features are listed in Table 3.1.

Algorithm	Assumptions	Assessment	Output	Characteristics
SGS [3]	I,II,III,IV	Added and omitted edges and arrows	MAG	Exponential search for conditional independence tests
PC [4]	I,II,III,IV	Added and omitted edges and arrows	MAG	Faster than SGS, independence tests conditional on all adjacent variables
FCI [5]	I,II,IV	PAG accuracy	PAG	Faster than PC, allows for latent variables
Any time FCI [6]	I,II,IV	PAG accuracy	PAG	Faster than FCI, outputs a correct although less informative structures
RFCI [7]	I,II,IV	PAG accuracy and run time	PAG	Faster than FCI, use fewer independence tests than FCI conditioned on less variables
Any-time FCI [8]	I,II,IV	PAG accuracy and run time	PAG	Can be interrupted anytime with true but incomplete information.
BCCD [9]	I,II,IV	PAG accuracy	PAG	Obtain a list of logical causal statements, orient edges by means of logical inference

Table 3.1: Main algorithms for learning causal structures from data. SGS stands for the Spirtes, Glymour and Scheines its inventors also PC stands for the initial of Peter Spirtes and Clark Glymour its inventors, FCI stands for Fast Causal Discovery, RFCI stands for Really Fast Causal Inference and finally BCCD stands for Bayesian Constraint-based Causal Discovery.

The early algorithms for causal discovery were aimed to solve the problem of determining a set of causal relations between variables and nothing else. For instance, SGS and PC algorithms are able to determine the structure of a PGM only when a large set of sample is at hand and assuming that every variable in the system is observed (measured) as well as there are not hidden common causes. This scenario totally lacks realism, as in many causal systems it is not always possible to gather the measurements for all variables and in others, one know that there may exist external factors which affect two variables thereby spurious relations can appear. In order to deal with scenarios more realistic FCI was proposed. By means of a PAG representation, FCI is able to find an equivalence class which bounds a causal structure close to the real one. However, FCI is able to do it in the large sample limit, whence a large number of sample is required. Thus far, the research was aimed to obtain a causal structure of the system although the computational effort was shelved. By changing the way of exploring the conditioning set of variables the proposed RFCI and Anytime FCI algorithms were aimed to deal with the complexity side. RFCI improves by reducing the required number of the statistical tests, and Anytime FCI attempts to obtain a true representation of the causal system but with less information with respect to the directionality of the links. In the other hand, BCCD is an attempt to hybridize the constraint-based and scores-based approaches by translating independence constraints into logic statements which allow mapping the independence relations to the presence or absence of causal relations. As the aim of BCCD algorithm is to improve the accuracy of the discovered relations, it is not able to deal with a dataset of a small number of samples. Thus, the proposed algorithm are aimed to consider the learning of the structure of a causal system in the presence of a small number of samples by means of the two goals, 1) the incorporation of prior knowledge, and 2) by combining the different data type as observational and experimental.

3.3 Combining observational and interventional data

Algorithms commented in Section 3.2 approach the recovery of a causal structure from a dataset of observations. This dataset is assumed sampled from the true original causal structure, however from observational data, the causal structure is only identifiable up to the same conditional independence relations often referred as the observational Markov equivalence class. Generally, interventional data improves the identifiability of the struc-

[3](Spirtes *et al.*, 2000)
[4](Spirtes *et al.*, 2000)
[5](Spirtes *et al.*, 2000)
[6](Spirtes, 2001)
[7](Colombo *et al.*, 2012)
[8](Spirtes, 2001)
[9](Claassen & Heskes, 2012)

ture (Hauser Alain, 2012), nevertheless, the extent of identifiability depends on the choice of the target variables. The area of modelling observational and interventional data has been vaguely explored. Next, the proposed works for address this problem are presented.

(Cooper & Yoo, 1999) proposed a Bayesian method for combining a mixture of observational and interventional data. Relying on causal sufficiency, a BN is learnt by assuming that the observational and experimental data is given at the beginning and the posterior probability was used as the evaluation metric. Similarly, Eaton and Murphy (Eaton & Murphy, 2007) applied the Koivisto and Sood dynamic programming algorithm (Koivisto & Sood, 2004) in order to compute the exact posterior probability from an experimental dataset. However, the problem of the identifiability of the Markov equivalence class is not addressed in their work. He and Geng proposed an active learning strategy for discovering causal structures (He Yang-Bo, 2008). First, from a set of observational data the Markov equivalence class is obtained. Later, the orientation of undirected edges is done by incorporating interventional data from randomized experiments and quasi-experiments. Although the algorithm can orient the edges of graph and output a DAG based on interventions, it needs to rely on the causal sufficiency assumption, i.e. assume that there are no latent variables. In 2007, Meganck *et al.* (Meganck *et al.*, 2007) proposed a method that compares two representation forms of the learnt causal model. In order to make the comparison, they scrutinized the Maximal Ancestral Graph and the Semi-Markovian representation given a series of advantages and disadvantages, as well as discover the network structure by means of an integral learning process from observational data and experimental data, after that they make some causal inference using both representations. Finally, they concluded that none of the existing techniques provide a complete answer to the problem of modelling systems with latent variables. Also in 2007, Borchani *et al.* Borchani *et al.* (2007) extends the Greedy Equivalence Search-Expectation Maximization algorithm to deal with an incomplete dataset of observational data joined with experimental data. In addition to this, proposed two strategies for selection of interventions; an adaptive and a non-adaptive approach. Finally, they compare both strategies respect to the total number of the performed interventions and the percentage of oriented edges.

Recently, Hauser and Bühlmann (Hauser & Bühlmann, 2014) proposed a Gaussian likelihood framework for joint modelling observational and interventional data. By means of Bayesian information criterion, the interventional Markov equivalence is estimated. The probability distribution of the observational dataset is assumed to be faithful (Markovian) to the true DAG structure, as well as the set of probability interventional distributions are linked to both the DAG and the observational distribution via *do-calculus*. Although the interventional Markov equivalence class is recovered, the underlying probability distribution needs to be Gaussian.

Causality has no a sharp definition, in fact, every approach discussed in Section 3.1 have emerged from different principles such as time series analysis, differential equations, probability distributions, among others. With the problem of the brain connectivity in mind, some researchers have extended existing techniques for causal discovery to more

realistic scenarios of brain connectivity, and others have defined new techniques entailing more biological characteristics. From a computational standpoint, the algorithms for causal discovery are as general as the kind of assumption allow them. However, as commented in Section 3.3 some researchers have realized of the necessity of collect (or generate) experimental data and to combine it along with observational data in order to improve the identifiability of the structures. Summarizing, there exist different works that are aimed to address the discovery of causal relation from data. Some of them have obtained promising results in the application to a problem domain.

Chapter 4

Research proposal

Most of the current techniques for learning the causal structure of a system are aimed to be able to deal with *all* kind of data. However, that is not always true since in order to acquire a reliable estimation of the true causal phenomena, one need to know several features about the data generation process. For instance, is data a multivariate dataset?, do we know about the existence of hidden variables? are the modelled variables the real variables in the system instead of a set of indirect observations? or even more important, are the events chronologically ordered?

It is true that a good solution must (or at least should) be a versatile solution, however in some application domains this is not always achieved. Since in order to make a set of causal claims about data, the data needs to rely on causal premises that cannot be inferred only from the data distribution but from a further knowledge as the answers to the previous questions. The aforementioned is summarized by Pearl as "behind any causal conclusion there must be some causal assumption" in (Pearl, 2009a).

Bearing this in mind, the characteristics of the input dataset often leads to a correct network, but this is not a sufficient condition to retrieve the maximal information (identify all the relevant features) in the connectivity network.

Following, the methodology to collect evidence to prove or refute the hypothesis stated in Section 1.5 will be described.

4.1 Methodology

In this work, a model for the causal discovery from data under the PGMs approach is investigated. A set of causal relations from multivariate random variables will be captured and subsequently the structure of the model will be enhanced in terms of completeness. The problem domain of the brain effective connectivity will require the capture of real data, consequently, a model for the discovery of causal (effective) relations between brain areas from fNIRS data is sought. The capture of these causal relations as well as the validation of the suggested model will be addressed considering the next

steps:

1. Develop a forward model which abides specific features according to system under study such as invariant links, common causes, among others properties in order to obtain synthetic data with predefined effective links.

Since it is desirable to have a well controlled scenarios which entails different causal features, it turns necessary the specification of a set of causal parameters into a forward model. This model shall embody predefined causal structures and predefined parameters in order to sample a dataset which will later be used for verification and validation purposes. The output of such a forward model mechanism is a structure network belonging to the space of solutions.

2. Identify the causal assumptions that can be relaxed with the aim of achieve a more informative causal structure regarding to the system under study.

Often, models rely on assumptions that are not always met in phenomena in nature. The different phenomena of brain function and fNIRS imaging are not the exception, so the assumptions presented in Section 2.2.4 will be scrutinised in order to check their compatibility with the dynamics of the brain and to understand the extent to which they can be relaxed.

3. Establish a procedure to incorporate a priori information about the underlying target structure into the stage of edges orientation in a causal discovery algorithm.

Most existing algorithms for causal structure learning work well under the assumption of a large sample availability. However, there are some application domains where this is not hold, (e.g. experiments in neuroscience). Consequently, a procedure must be proposed to alleviate undefined relations often yielded by small samples. In this research, it is hypothesized that the addition of a priori information is an appropriate candidate to address this issue.

4. Identify a plausible set of target variables over which to perform theoretical interventions.

Given the information obtained in stage 2, a strategy to identify a set of target variables over which to perform a disturbance. In order to identify the set of potential variables to disturb, some aspects of the domain knowledge have to be considered. Thus, the basic idea is to identify the variables which can be ideally[1] intervened. This disturbance will be computed by means of the *do-calculus*.

[1]An intervention on a variable is said *ideal* when it is possible to isolate the variable from direct causes and also the effect of such an intervention only affects the probability distribution of its direct effects.

5. Define a strategy for combining observational source data and new generated interventional (experimental) data in order to improve the completeness of the network discovered.

The *do* operator will be applied in order to compute the causal effects (i) on the child variables (direct effects) and (ii) on the complete network, so it will be possible to know how the joint probability distribution change. Once the target variables are selected, it shall be possible to sample the new joint probability distribution in order to combine it with previous samples. This new data set is expected to be used to improve the network by unravelling the undefined directionality of some links as well as identify some spurious of them.

6. Incorporate the previous elements in order to develop a working algorithm for causal PGM discovery.

As a result of above stages, a working algorithm can be developed. In this case, it requires the assumptions on which it relies, besides the input of a priori information in the form of constraints, as well as identify the set of target variables to latter perform a series of interventions. Assuming a change in the joint probability distribution, the algorithm will generate a sample (interventional data) which jointly with the original observational sample will be combined to distil how much information about causal directions can be further retrieved. At the end, an equivalence class with less uninformative features (undefined links) is expected as the output of the algorithm.

7. Verification of the algorithm proposed in stage 6 with synthetic dataset generated by a known model as well as using a benchmarking set.

Considering the model stated in stage 1, this will be tuned to have specific causal patterns in the connectivity it express, so the algorithm has to be able to recover some of these patterns. It is expected that those relations not identified at the beginning, will be captured when the results of interventions are combined with the source of observations.

8. Design an fNIRS experiment involving cognitive or sensor tasks to obtain domain specific data.

An experiment for capturing the underlying brain effective connectivity network will be performed in order to collecting a dataset using the continuous wave fNIRS neuroimaging modality. The design of the experiment demands the choice of a cognitive or motor task. The chosen task must involve a well-known neural circuit, so that the proposed model is able to recover the connections between brain regions previous reported in the literature. This can be partially tested with standard segregation analysis, since data will be collected by means of a continuous wave fNIRS device, the circuit of choice must be confined to the cortical region due to the limited depth reached by this neuroimaging modality.

9. Collecting data from fNIRS cognitive or sensor experiments.

The data collection, will be carried out during a secondment at the Biomedical Optics Research Laboratory under the supervision of Ilias Tachtsidis at the University College London. Although the secondment is mainly focused at collecting data for brain connectivity, also is expected to investigate the causal relations among the neural responses and systemic interference in fNIRS.

10. Validate the algorithm for causal discovery.

Using the proposed algorithm for causal discovery at stage 6 and the dataset from an fNIRS experiment collected at the stage 9, the circuit of brain effective connectivity will be recovered. The aim is to nomologically validate the proposed model as capable of recover the connections between brain regions of the known circuit previously reported in the literature.

Table 4.1 shows the relation between stages in the methodology, objectives and research questions.

Stage	**Task**	**Specific objective**	**RQ**
1	Development of a forward model with a set of links constraints.	6	1
2	Identify the causal assumptions that can be relaxed.	3	2
3	Incorporate a priori information of the structure problem at hand.	2	2
4	Identify a set of target variables to perform interventions.	4	3
5	Combining observational along with interventional data.	5	3
6	Development of a working algorithm for causal PGM discovery.	1	1
7	Verification of the proposed algorithm with synthetic data.	6	2
8	Design an fNIRS experiment involving cognitive task.	6	1
9	Collecting data from designed fNIRS experiments.	6	1
10	Validate the causal discovery algorithm.	6	1

Table 4.1: Relation between task in the methodology (Sec. 4.1) and its objectives (Sec. 1.6)and research questions (RQ) associated (Sec. 1.4).

4.2 Publications plan

The publications plan shown in Table 4.2 will be attempted:

Name	Type	IF	Contribution	RQ	OBJ	Submission
FLAIRS	C	-	Efficient mechanism to incorporate a priori information in a causal discovery algorithm.	2	2	Nov 2015
PGM	C	-	A method to select suitable target variables over which performing experimental disturbances.	3	4	May 2016
UAI	C	-	A mechanism to join interventional along with observational data elucidating the directionality in PGMs.	3	5	Mar 2017
Neuroimage	J	6.3	A causal PGM for the analysis of the effective connectivity in fNIRS.	1	6	July 2017
AI	J	3.3	Algorithm for causal discovery alleviating classical assumptions.	1	1	Dec 2017

Table 4.2: Publications plan, including journals (J) and conference (C) papers, impact factor (IF), intended submission date and the relation with the research questions (RQ) and objectives (OBJ). AI: Artificial Intelligence, FLAIRS: The Florida Artificial Intelligence Research Society, PGM: Probabilistic Graphical Models, UAI: Uncertainty in Artificial Intelligence.

4.3 Schedule of activities

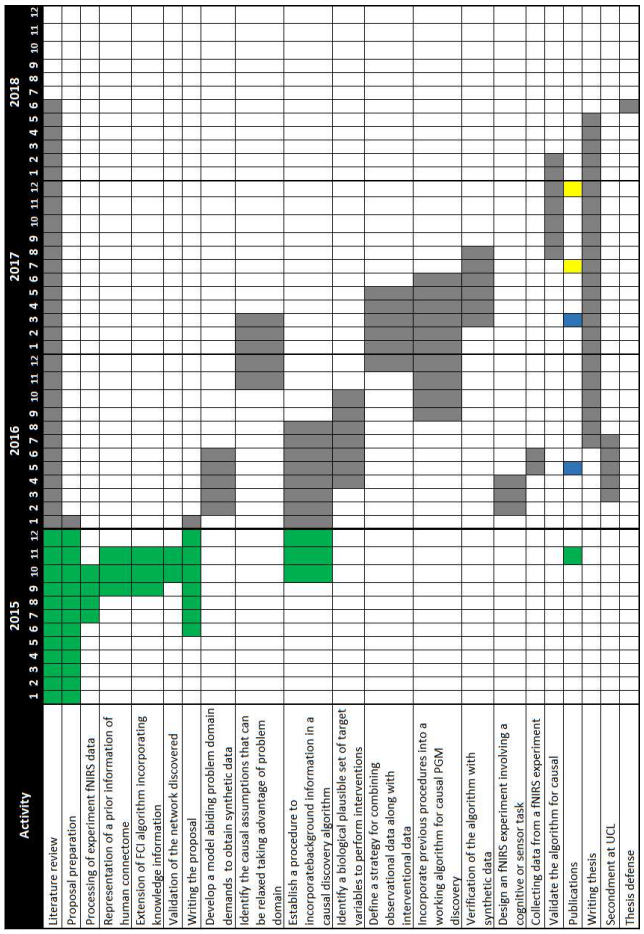


Figure 4.1: Schedule of activities per month for the PhD period. Green cells indicate activities acomplished, cells in gray indicate future activities, yellow cells indicates journal publications and blue cells indicate conference publications.

Chapter 5

Preliminary Results

In this first year, some preliminary aspects of the research problem have already been addressed. Specifically, the specification of a priori information about some neural circuits and its addition to an extension of the FCI causal algorithm. Next, the newly proposed seed FCI (sFCI) algorithm, an extension to the classical FCI is presented, as well as the methods and the results of the first experiment carried out, summarizing the progress in the first year.

5.1 Seeded FCI

In some situations, there might be the possibility of incorporating prior information to the model in order to elucidate undefined relations. The addition of a priori information may resolve some of the undecided directions present in the output of the FCI algorithm i.e. the partial ancestral graph. This prior knowledge can be added in form of restrictions either by defining fundamental or unnecessary relations. Incorporating a priori knowledge to resolve undefined relations requires the modification of the basic FCI algorithm. Algorithm 1 is the current proposal, a seeded version of FCI (sFCI), for permitting incorporation of prior information to FCI.

The basic idea in sFCI is to start with a complete undirected graph $\mathcal{Q}(\mathbf{V}, \mathbf{E})$ and a set of invariant links $\mathbf{L}$ (prior information). Then iteratively, select a pair of adjacent variables X, Y in $\mathcal{Q}$ and select a subset of adjacent variables to both X and Y, the link between X and Y is removed if they are d-separated and they are not in $\mathbf{L}$, otherwise hold it, and so on for the rest of adjacent pairs. Next, all edges are oriented as undefined ($\circ$) in both extremes and using the result of d-separation test it is possible to reorient the triplets of the form $A * - * B * - * C$.

The structural connections of the human brain establishes a set of anatomical constraints with respect to the possible paths in its connectivity. In neuroimaging, this set of constraints can be obtained from the so called *human connectome* (Hagmann *et al.*, 2008; Joshi *et al.*, 2010), which establishes the expected physical links in the human

brain.

Algorithm 1: Seeded Fast Causal Inference (sFCI) algorithm with a mechanism to consider prior information. The ariables $n, \mathbf{S}, \mathbf{V}, \mathbf{E}, \mathbf{L}$ are the number of conditioning variables, the set of conditioning variables, the set of variables, the set of edges and the set of *a priori* links respectively. The processes Adjacencies$(\mathcal{Q}, X)$, d-separated$(X, Y|\mathbf{S})$, Possible-D-SEP(B, A) returns the set of adjacencies of node X in graph $\mathcal{Q}$, "true" if the variables X, Y are d-separated given the set $\mathbf{S}$, the set of nodes which possibly d-separate variable A and B, respectively.

Data: set of variables $\mathbf{V}$, set of a priori links $\mathbf{L}$
Result: A partial ancestral graph $\mathcal{F}$

1) Start a complete undirected graph $\mathcal{Q}(\mathbf{V}, \mathbf{E})$ over the set of nodes $\mathbf{V}$;
2) $n = 0$;
repeat
  repeat
    select an ordered pair $X - Y \in \mathbf{E}$ such that $|\text{Adjacencies}(\mathcal{Q}, X)\backslash\{Y\}| \geq n$, and a subset $\mathbf{S} \subseteq \text{Adjacencies}(\mathcal{Q}, X)\backslash\{Y\}$ such that $|\mathbf{S}| = n$;
    if *d-separated*$(X, Y|\mathbf{S})$ *and* $X - Y \notin \mathbf{L}$ **then**
      delete the edge $X - Y$ from $\mathbf{E}$;
      record $\mathbf{S}$ in Sepset(X, Y) and Sepset(Y, X);
  until $\forall\{X, Y\} : |$*Adjacencies*$(\mathcal{Q}, X) \backslash\{Y\}| \geq n$ *and* $\forall \mathbf{S} \subseteq$ *Adjacencies*$(\mathcal{Q}, X)\backslash\{Y\}$ *such that* $|\mathbf{S}| = n$ *have been tested for d-separation* ;
  $n = n + 1$;
until $\forall\{X, Y\} : |$*Adjacencies*$(\mathcal{Q}, X) \backslash\{Y\}| < n$;
3) Let $\mathcal{F}(\mathbf{V}, \mathbf{E}')$ be the undirected graph resulting from step 2), then orient each edge as $\circ - \circ$;
$\forall A - B - C$ such that $A - B, B - C \in \mathbf{E}'$ and $A - C \notin \mathbf{E}'$ and $A - B, B - C \notin \mathbf{L}$;
  if $B \notin$ *Sepset*(A, C) **then**
    orient $A * - * B * - * C$ as $A * - > B < - * C$
4) $\forall A - B \in \mathbf{E}'$ and $A - B \notin \mathbf{L}$;
  if *d-separated*$(A, B|\mathbf{S})$ *such that* $\mathbf{S} \in$ *Possible-D-SEP*$(A, B)\backslash\{A, B\}$ *or Possible-D-SEP*$(B, A)\backslash\{A, B\}$ **then**
    remove $A - B$;
    record $\mathbf{S}$ in Sepset(A, B) and Sepset(B, A);

5.2 Decoding effective connectivity in fNIRS

Unlike the functional connectivity, effective connectivity is concerned with decoding coordinated action of brain regions, and most importantly, determining the direction of the flow of information. The brain regions can be modelled as variables within a system and the coordination between them as independence relations. Considering the spatial resolution of common fNIRS systems[1], it is possible to assume that each channel interrogates a certain brain region, and that the cortical area interrogated by a certain channel does not overlap with that of other channels.

[1]State of the art high definition diffuse optical tomography (HD-DOT) will invalidate this assumption (Eggebrecht *et al.*, 2014)

In this sense, the system of interest is a fNIRS neuroimaging which is able to take a snapshot of the cortical activity across brain regions while the subject is performing a certain task. During the fNIRS image acquisition, a number of observations of the system can be gathered corresponding to task repetitions, i.e. blocks or trials. A critical limitation in the fNIRS domain is that this set of observations only can incorporate a reduced number of samples due to the *habituation* effect in the brain, a response decrement due to stimulus repetition (Fischer *et al.*, 2000), effectively limiting the number of trials per session.

In addition, although the effective neural circuits are dynamic, the brain structural connectivity establishes a set of restrictions regarding to their anatomical paths. The full set of structural neural pathways is described by the *connectome* (Hagmann *et al.*, 2008; Joshi *et al.*, 2010).

With this in mind, it is possible to state the problem of recovering the effective connectivity evoked by a certain task as observed with fNIRS as the modelling of causal relations among a set of variables considering a limited set of samples and integrating prior information e.g. from the connectome.

5.3 Experiment

The fNIRS neuroimaging dataset for this research was originally collected at Imperial College London back in 2007 to question about experience-dependent differences on prefrontal activity for a cohort of surgeons while knot-tying (Leff *et al.*, 2007). Details of the data collection can be found in the original publication, but briefly 62 surgeons (19 consultants, 21 trainees and 22 medical students) participated in the study[2]. Brain activity was monitored with a 24-channel fNIRS system (ETG-4000, Hitachi Medical Co., Japan). All channels were set in the prefrontal cortex, with inferior channels 2 and 13 close to F8 and F7 international 10/10 locations respectively and superior channels 12 and 23 close to FC4 and FC3 respectively. The surgical knot-tying task was repeated 5 times at self-pace, allowing 30 seconds recovery between trials. Data was collected at 10Hz and changes in concentration of oxygenated and reduced haemoglobin reconstructed using the modified Beer-Lambert law.

For this exercise, each optical neuroimage was detrended and decimated to 1Hz in order to remove the system drift as well as unrelated physiological signals. Data was split by blocks and later resampled to obtain blocks of equal length, equal to the mean trial duration for each expertise group. Finally, data was group averaged across subjects within each group in order to increase signal to noise ratio (SNR) and obtain a common structure for each group. For this particular exercise, only oxygenated haemoglobin HbO_2 often considered to have a more favourable SNR of the two species (Orihuela-Espina *et al.*, 2010) was used. The connectome information was recovered from (Hagmann *et al.*, 2008;

[2]The authors kindly allowed the use of the dataset for this research

Joshi *et al.*, 2010) and adapted to the channel location described in (Leff *et al.*, 2007), and it is shown in Fig 5.1.

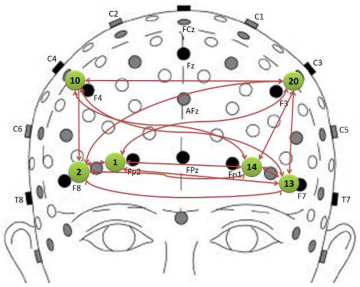


Figure 5.1: Structural information from the connectome used as prior knowledge (Hagmann *et al.*, 2008; Joshi *et al.*, 2010).

5.4 Results

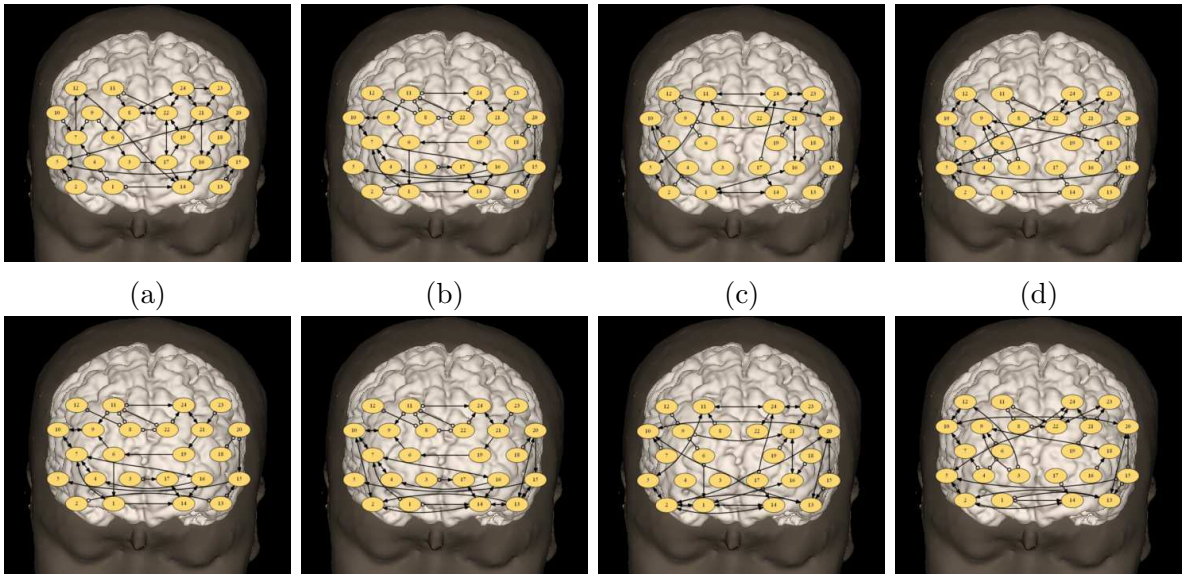
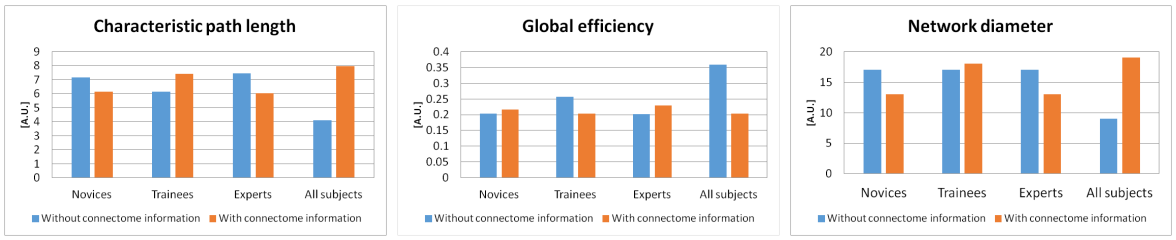


Figure 5.2: Effective connectivity networks. Top row show the connectivity revealed with plain FCI whereas bottom row show the connectivity networks exploiting prior information from the connectome with sFCI. Columns represent the different expertise groups; from top to bottom; novices (a and e), trainees (b and f), and consultants (c and g) and all-subjects (d and h) networks.

A total of eight networks presented in Fig. 5.2 were built considering four groups (novices, trainees, experts and all subjects) and two variants, with and without connectome information using the sFCI in Algorithm 1 and "classical" FCI respectively. Table 5.1 summarizes the number of undefined links using both FCI and our proposal sFCI. As expected, the number of undefined links decreases with the utilization of prior information.

Algorithm	Novices	Trainees	Experts	All subjects
FCI	11	19	18	26
sFCI	8	16	14	21

Table 5.1: Number of undefined links in the networks for each group using FCI and sFCI algorithms.



(a) (b) (c)

Figure 5.3: Measures of brain connectivity. (a) Characteristic path length, (b) Global efficiency and (c) Network diameter. The characteristic path length represents the average shortest path length between each pair of nodes in a network (Watts & Strogatz, 1998). The global efficiency measures the average inverse shortest path in the network and its proportional to the brain capacity of parallel information flow (Latora & Marchiori, 2001). The network diameter is the largest number of nodes in order to travel from one node to another when loops are not considered (West *et al.*, 2001).

Fig. 5.3 shows graph-theoretical measures of functional integration. When the information of the connectome is not considered, our results clearly align with previous evidence. In the work where the dataset was originally collected (Leff *et al.*, 2007) the greater activity of the novices' prefrontal cortex leaded to a more lateralised response; an effect which can also be appreciated in the effective networks in Fig. 5.2. Interestingly, the network metrics in Fig. 5.3 tell a story very much alike that found in (Ohuchida *et al.*, 2009) in which trainees evoked higher activity than novices or experts. These two observations provides preliminary nomological evidence about the working of the model. When the prior information is considered, the network features tell a different story

regarding the configuration of the network e.g. different pattern of path length and diameter, but shows a more intuitive higher efficiency of the network in experts, suggesting that the incorporation of the expected structural information can perhaps reveal more realistic effective information. However, further evidence is need before such claim can be made.

5.5 Conclusions

Motivated by the neuroscientific demand of better modelling tools for the analysis of effective connectivity, this thesis proposes the generation of a causal PGM by means of an algorithm for causal discovery. In turn, three major issues related to the learning of causal structures will be addressed. The first issue is to learn the PGM structure with only a limited set of samples. Secondly, it is proposed to develop a mechanism in order to select a subset of variables over which ideal interventions can take place. Thirdly, the incorporation of background information in the model have been proposed in order to partially circumvent the small samples limitation. Finally, the combination of observed variables jointly with disturbed variables is proposed in order to improve the number of links corrected o missed.

Thus far, a first experiment extending the FCI algorithm to be able to take the advantage of the background (*a priori*) information was performed. The addition of the background information was extracted from the connectome of the human brain. The performance of the proposed algorithm was assessed in terms of the number of invariant marks present in the model generated. The result of the connectivity network (See 5.4) was consistent with others works in the literature strengthening this proposal in terms of the feasibility of the second objective (See Section 1.6).

Finally, it is worth to highlighting that the direction of the research seems already partially adequate and the proposal appears feasible with one of the specific objectives have been initially addressed experimentally with the obtained results being promising.

Bibliography

Pieter Abbeel, Daphne Koller, & Andrew Y Ng. Learning factor graphs in polynomial time and sample complexity. *The Journal of Machine Learning Research*, 7:1743–1788, 2006.

P. Achermann & A. A. Borbély. Coherence analysis of the human sleep eclectroencephalogram. *NeuroScience*, 85(4):1195–1208, 1998.

David A Boas, Constantinos Pitris, & Nimmi Ramanujam. *Handbook of biomedical optics*. CRC press, 2011.

Hanen Borchani, Maher Chaouachi, & Nahla Ben Amor. Learning Causal Bayesian Networks from Incomplete Observational Data and Interventions, 2007.

David Maxwell Chickering. Learning equivalence classes of bayesian-network structures. *The Journal of Machine Learning Research*, 2:445–498, 2002.

Thomas Claassen. *Causal discovery and logic*. PhD thesis, Radboud University Nijmegen, 2013.

Tom Claassen & Tom Heskes. A Bayesian Approach to Constraint Based Causal Inference. *UAI 2012, Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*, pages 207–216, 2012.

Diego Colombo, Marloes H. Maathuis, Markus Kalisch, & Thomas S. Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, 40(1):294–321, 2012.

G Cooper & C Yoo. Causal discovery from a mixture of experimental and observational data. *Proc. Fifthteenth Conference on Uncertainty in Artificial Intelligence {(UAI'99)}*, pages 116–125, 1999.

Mark Cope, David T. Delpy, E. O. R. Reynolds, Susan Wray, J. S. Wyatt, & P. Zee van der. Methods of quantitating cerebral near infrared spectroscopy data. *Advances in Experimental Medicine and Biology*, 222:183–189, 1988.

Marek J Druzdzel. The role of assumptions in causal discovery. *UNCERTAINTY PROCESSING*, page 57, 2009.

Daniel Eaton & Kevin P. Murphy. Exact Bayesian structure learning from uncertain interventions. *International Conference on Artificial Intelligence and Statistics*, 2007.

Frederick Eberhardt. *Causation and Intervention*. PhD thesis, Carnegie Mellon University, 2007.

Adam T Eggebrecht, Silvina L Ferradal, Amy Robichaux-Viehoever, Mahlega S Hassanpour, Hamid Dehghani, Abraham Z Snyder, Tamara Hershey, & Joseph P Culver. Mapping distributed brain function and networks with diffuse optical tomography. *Nature photonics*, 8(6):448–454, 2014.

Håkan Fischer, Tomas Furmark, Gustav Wik, & Mats Fredrikson. Brain representation of habituation to repeated complex visual stimulation studied with pet. *Neuroreport*, 11(1):123–126, 2000.

Richard SJ Frackowiak, Karl J Friston, Christopher D Frith, Raymond J Dolan, Cathy J Price, Semir Zeki, John T Ashburner, & William D Penny. *Human Brain Function*. Academic Press, 2004.

Karl J. Friston. Functional and Effective Connectivity: A Review. *Brain Connectivity*, 1(1):13–36, 2011.

Karl J. Friston, Chris D. Frith, P. Fletcher, P. F. Liddle, & Richard S. J. Frackowiak. Functional topography: Multidimensional scaling and functional connectivity in the brain. *Cerebral Cortex*, 6:156–164, 1996.

Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.

Russell Greiner, Adam J. Grove, & Dale Schuurmans. Learning bayesian nets that perform well. *CoRR*, abs/1302.1542, 2013.

Patric Hagmann, Leila Cammoun, Xavier Gigandet, Reto Meuli, Christopher J Honey, Van J Wedeen, & Olaf Sporns. Mapping the structural core of human cerebral cortex. *PLoS Biol*, 6(7):e159, 2008.

Alain Hauser & Peter Bühlmann. Two optimal strategies for active learning of causal models from interventional data. *International Journal of Approximate Reasoning*, 55 (4):926–939, jun 2014.

Bühlmann Peter Hauser Alain. Two optimal strategies for active learning of causal modeles form interventios. In *Probabilistic Graphical Models*, 2012.

Geng Zhi He Yang-Bo. Active Learning of Causal Networks with Intervention Experiments and Optimal Designs. *Journal of Machine Learning Research*, 9:2523–2547, 2008.

Rick H Hoyle. *Handbook of structural equation modeling*. Guilford Press, 2012.

A.A. Joshi, S.H. Joshi, I. Dinov, D.W. Shattuck, R.M. Leahy, & A.W. Toga. Anatomical structural network analysis of human brain using partial correlations of gray matter volumes. In *Biomedical Imaging: From Nano to Macro, 2010 IEEE International Symposium on*, pages 844–847, April 2010.

Markus Kalisch & Peter Bühlmann. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *The Journal of Machine Learning Research*, 8:613–636, 2007.

Markus Kalisch, Martin Machler, Diego Colombo, Marloes H Maathuis, & Peter Buhlmann. Causal Inference Using Graphical Models with the R Package pcalg. *Journal of Statistical Software*, 47(11):26, 2012.

Stefan J Kiebel, Marta I Garrido, Rosalyn Moran, Chun-Chuan Chen, & Karl J Friston. Dynamic causal modeling for eeg and meg. *Human brain mapping*, 30(6):1866–1876, 2009.

Mikko Koivisto & Kismat Sood. Exact bayesian structure discovery in bayesian networks. *The Journal of Machine Learning Research*, 5:549–573, 2004.

Daphne Koller & Friedman Nir. *Probabilistic Graphical Models, principles and techniques*. The MIT Press, 2009. doi: 10.1007/s13398-014-0173-7.2.

S. Kullback & R. A. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22(1): 79–86, 03 1951.

Vito Latora & Massimo Marchiori. Efficient behavior of small-world networks. *Physical review letters*, 87(19):198701, 2001.

Daniel Richard Leff, Felipe Orihuela-Espina, Louis Atallah, Ara W. Darzi, & Guang-Zhong Yang. Functional near infrared spectroscopy in novice and expert surgeons: a manifold embedding approach. In N. Ayache, S. Ourselin, & A. Maeder, editors, *Medical Image Computing and Computer-Assisted Intervention (MICCAI'07)*, volume 4792, pages 270–277, Australia, 2007. Lecture Notes in Computer Science.

Junning Li, Z JaneWang, & Martin J McKeown. Graphical models of functional mri data for assessing brain connectivity. In *Neuroimaging*, pages 375–396. InTech, 2012.

Chris Martin. Contributions and complexities from the use of in-vivo animal models to improve understanding of human neuroimaging signals. *Frontiers in Neuroscience*, 8 (211), 2014.

Stijn Meganck, Philippe Leray, & Bernard Manderick. *Causal Graphical Models with Latent Variables: Learning and Inference*, pages 5–16. Springer Berlin Heidelberg, 2007.

Richard E. Neapolitan. *Learning Bayesian Networks*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2003. ISBN 0130125342.

Kenoki Ohuchida, Hajime Kenmotsu, Atsuyuki Yamamoto, Kazuya Sawada, Takehito Hayami, Kenichi Morooka, Shinichiro Takasugi, Kozo Konishi, Satoshi Ieri, Kazuo Tanoue, Yukihide Iwamoto, Masao Tanaka, & Makoto Hashizume. The frontal cortex is activated during learning of endoscopic procedures. *Surgical Endoscopy*, page In press, 2009.

F Orihuela-Espina, DR Leff, DRC James, AW Darzi, & GZ Yang. Quality control and assurance in functional near infrared spectroscopy (fnirs) experimentation. *Physics in medicine and biology*, 55(13):3701, 2010.

Judea Pearl. *Probabilistic reasoning in intelligent systems*. Morgan Kaufmann, 1988.

Judea Pearl. *Causality: models, reasoning and inference*, volume 29. Cambridge Univ Press, first edition, 2000.

Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146, 2009a.

Judea Pearl. *Causality: models, reasoning and inference*. Cambridge Univ Press, second edition, 2009b.

Jagath C. Rajapakse & Juan Zhou. Learning effective brain connectivity with dynamic bayesian networks. *NeuroImage*, 37(3):749 – 760, 2007.

Thomas Richardson & Peter Spirtes. Ancestral graph Markov models. *Annals of Statistics*, 30(4):962–1030, 2002.

R W Robinson. Counting labeled acyclic digraphs. *New Directions in Graph Theory*, pages 239–273, 1973. doi: 10.1007/bfb0069178.

R W Robinson. Counting unlabeled acyclic digraphs. *Combinatorial Mathematics V*, 622:28–43, 1977. doi: 10.1007/bfb0069178.

Bruce Rosen, Van J Wedeen, JDV Horn, Bruce Fischl, Randy L Buckner, Lawrence Wald, Matti Hamalainen, Steven Stufflebeam, Joshua Roffman, David W Shattuck, *et al.* The human connectome project. In *Organization for Human Brain Mapping Annual Meeting. Barcelona, Spain*, 2010.

Richard Scheines. An Introduction to Causal Inference. *Causality in crisis?*, 1997.

R.E. Schumacker & R.G. Lomax. *A Beginner's Guide to Structural Equation Modeling*. Lawrence Erlbaum Associates, 2004.

Peter Spirtes. An anytime algorithm for causal inference. In *Proc. of the Eighth International Workshop on Artificial Intelligence and Statistics*, pages 213–221, 2001.

Peter Spirtes. Introduction to causal inference. *J. Mach. Learn. Res.*, 11:1643–1662, August 2010.

Peter Spirtes, Clark Glymour, & Richard Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, Massachusetts, USA, 2nd editio edition, 2000.

Olaf Sporns, Giulio Tononi, & Rolf Kötter. The human connectome: A structural description of the human brain. *PLoS computational biology*, 1, 2005.

Gary Strangman, David A Boas, & Jeffrey P Sutton. Non-invasive neuroimaging using near-infrared light. *Biological Psychiatry*, 52(7):679 – 693, 2002.

L.E. Sucar. *Probabilistic Graphical Models: Principles and Applications*. Advances in Computer Vision and Pattern Recognition. Springer London, 2015.

Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Working Group. Brain 2025 a scientific vision. Technical report, National Institute of Health, June 2014.

Arno Villringer & Britton Chance. Non-invasive optical spectroscopy and imaging of human brain function. *Trends in neurosciences*, 20(10):435–442, 1997.

Duncan J Watts & Steven H Strogatz. Collective dynamics of small-world networks. *Nature*, 393(6684):440–442, 1998.

Douglas Brent West *et al. Introduction to graph theory*, volume 2. Prentice hall Upper Saddle River, 2001.

Sewall Wright. Correlation and causation. *Journal of agricultural research*, 20(7):557–585, 1921.

Or Zuk, Shiri Margel, & Eytan Domany. On the number of samples needed to learn the correct structure of a bayesian network. *arXiv preprint arXiv:1206.6862*, 2012.