Knowledge Transfer for Learning Subject-Specific Causal Probabilistic Graphical Models

**Verónica Rodríguez López, Luis Enrique Sucar Succar,
Felipe Orihuela Espina**

Technical Report No. CCC-19-004
November 2019

Luis Enrique Erro 1
Sta. Ma. Tonantzintla,
72840, Puebla, México

Abstract

There are domains, such as biology, medicine, and neuroscience, where the causal relations vary across members of a population, and where it may be difficult to collect data for some specific members. For these domains, it is convenient to develop algorithms to learn subject-specific causal models. Causal probabilistic graphical models have shown to be a tool for modeling probabilistic causal relations. Most of the algorithms for learning causal graphical probabilistic models are inadequate for learning subject-specific models, especially for subjects with a limited dataset, since they were designed to find the common causal relations of a population in the large sample limit. Although there are algorithms for partially learning subject-specific causal models, they are limited for learning only from observational datasets. The main goal of this research is to develop a knowledge transfer algorithm for learning subject-specific causal probabilistic graphical models. We expected to contribute with an algorithm that, transferring observational and interventional data, together with causal relations of related sources, learns the structure of causal probabilistic graphical models. We hypothesized that leveraging knowledge from auxiliary sources may help to reliably identify the special causal relations of specific subjects with a limited dataset. Synthetic causal probabilistic graphical models together with interventional and observational samples from them will be used to validate our algorithm. Besides, we will plan to exemplify our algorithm in a causal analysis problem of the neuroscience domain. Preliminary results about the feasibility of extending a score-based discovery causal algorithm with transfer learning techniques, and also related to an experiment of functional brain connectivity analysis using probabilistic graphical models are presented.

Keywords

Causal discovery, Probabilistic graphical models, Subject-specific causal models

Contents

Acronyms

CMC Causal Markov Condition

CBN Causal Bayesian Network

CS Casual Sufficiency

DAG Directed Acyclic Graph

FC Faithfulness Condition

FCI Fast Causal Inference

FGES Fast Greedy Equivalence Search

GES Greedy Equivalence Search

GFCI Greedy Fast Causal Inference

GIES Greedy Interventional Equivalence Search

IGES Instance specific Greedy Equivalence Search

MEC Markov Equivalence Class

PDAG Partially Directed Acyclic Graph

PGM Probabilistic Graphical Model

NSHD Normalized Structural Hamming Distance

Notation

X : A random variable.

x : A value of random variable X.

$\mathbf{X}$: A set of random variables, $\mathbf{X} = \{X_1, X_2, ..., X_N\}$.

$\mathbf{x}$: An assignment of value to each random variable in a set $\mathbf{X}$.

$\mathbf{Pa}(X)$: Parents of node X in a directed graph.

$\mathbf{pa}(X)$: Values of the parents of node X in a directed graph.

$\mathcal{X}$: Space of features in Transfer Learning.

$\mathcal{Y}$: Space of labels in Transfer Learning.

$\mathcal{D}$: Domain in Transfer Learning.

$\mathcal{T}$: A task in Transfer Learning.

f($\cdot$) : A function.

Chapter 1

Introduction

Causation is a relation between two particular events where there is an event that affects the other (Spirtes *et al.*, 2000). It is an important concept for several domains where it is necessary to understand the process of data generation and infer effects of manipulations over some elements of a system. In particular, causal probabilistic graphical models (causal PGMs) are useful tools for these domains, since they encode causal relations between the variables of systems and provide information to make predictions under manipulations (Heinze-Deml *et al.*, 2018). In their learning, two aspects are important: causal structure learning and parameters estimation from data and the causal structure. Causal structure learning consists in discovering the causal relations between the variables of a system either from observations, throught interventions or both, with the first one being inherently limited. Parameter estimation refers to estimate the joint probabilistic distribution of the model.

In the learning of causal models, randomized experiments should be used since they provide information about the effects of certain manipulations over a system. A combination of these experimental data, together with the measurement obtained from the variables of a system without performing manipulations (referred to as observational data), allows reliable learning the structure of causal probabilistic graphical models (causal PGMs).

Some domains, such as biology, neuroscience, and medicine, where the causal relations vary across members of a population (Cooper *et al.*, 2018; Mechelli *et al.*, 2002; Monleón Getino & Canela i Soler, 2017), demand the learning of subject-specific causal PGMs that encode the specific causal relations for a particular member of a population. Although there are many studies in the field of causal PGMs that have explored the structure learning of causal models, most of these studies are focused on finding population-wide causal PGMs that encode the common causal relations of a population (Glymour *et al.*, 2019; Malinsky & Danks, 2018; Mooij *et al.*, 2019; Spirtes & Zhang, 2016; Tillman & Eberhardt, 2014). Only a few works have addressed the problem of learning subject-specific causal PGMs (Jabbari *et al.*, 2018; Jia *et al.*, 2018; Li *et al.*, 2018).

Learning subject-specific causal PGMs using the existing discovery causal algorithms imposes some difficulties. Most of these algorithms find the true causal structure in the large sample limit (Mooij *et al.*, 2019; Ogarrio *et al.*, 2016; Tillman & Eberhardt, 2014; Zhang *et al.*, 2018). However, because of the physical condition of the subjects, the difficulty or cost to carry out experiments, it can be complicated collecting enough data for the learning, especially for some subjects. Moreover, traditional algorithms, such as PC (Spirtes & Glymour, 1991) and GES (Chickering, 2002), partially find the common casual relations of a population from a single observational dataset. The heterogeneity of data due to differences in sampling methods and experimental conditions is not considered on these algorithms. Although there are specialized algorithms that combine multiples dataset considering their heterogeneity, they were also designed to find population-wide causal PGMs (Claassen & Heskes, 2010; Hauser & Bühlmann, 2012; Mooij *et al.*, 2019; Ramsey *et al.*, 2010; Tillman & Spirtes, 2011; Triantafillou & Tsamardinos, 2015). Recently, more appropriate algorithms to learn subject-specific causal PGMs have been developed (Jabbari *et al.*, 2018; Jia *et al.*, 2018; Li *et al.*, 2018). Some of these algorithms assume that there is a sufficient number of samples (Zuk *et al.*, 2012) for the learning (Jabbari *et al.*, 2018; Li *et al.*, 2018), limiting the learning to observational data (Jabbari *et al.*, 2018; Jia *et al.*, 2018), or assuming homogeneity in causal relations with only variations in causal effects across subjects (Li *et al.*, 2018).

Due to the difficulties of existing discovery causal algorithms, mainly regarding for discover subject-specific causal PGMs from a limited dataset, the aim of this research is the development of an algorithm for structure learning of subject-specific causal PGMs, that use the knowledge of datasets composed by experimental and observational samples and auxiliary causal PGMs. In this document, the research proposal related to this transfer knowledge algorithm is described.

1.1 Motivation

In several domains, there are specific causal relations for some member of a population. For example in neuroscience, because of differences in the degree of disease affectation and the recovery process, it has been observed that causal relations between brain regions might vary across patients (Grefkes & Fink, 2014; Li *et al.*, 2008; Mechelli *et al.*, 2002; Wu *et al.*, 2011). Studies in genetic and medicine have also suggested variations in causal relations across subjects. Findings in genetics have revealed that there are somatic genome alterations causing expression changes in specific tumors (Cooper *et al.*, 2018). While in medicine, it has been observed, for reasons of genetic, and environmental factors or disease stage, that the effect of drugs could vary across patients (Monleón Getino & Canela i Soler, 2017).

In these domains where there are variations on causal relations across subjects, subject-specific causal models are suitable for understanding the generation process in specific members of a population. These models could help to capture the specific causal relations of a particular subject at some stage of interest, such as disease or recovery stage.

1.2 Justification

This research will address the following reported open problems for the area of causal probabilistic graphical models:

Limited sample sizes (Spirtes & Zhang, 2016): This problem refers to how to correct the lack of enough data for learning causal PGMs. Most of the existing discovery causal algorithms reliably find causal models when there is a sufficient number of samples (Zuk *et al.*, 2012). In some situations, collecting enough data for some specific members of a population could be difficult. Transferring instances of auxiliary datasets, or causal relations from auxiliar causal PGMs, could compensate for lack of enough data and could allow learning more accurate causal models for specific subjects.

Heterogeneity of data (Bareinboim & Pearl, 2016; Glymour *et al.*, 2019): This issue refers to finding the appropriate causal PGMs that best adjust to a combination of multiples experimental and observational datasets, considering that they were collected from diverse populations, or under different experimental conditions and sampling methods.

1.3 Problem Statement

There are several issues to be considered in the structure learning of subject-specific causal PGMs. The first one is the lack of enough data for target subject, that we will try to compensate in this research transferring instances of auxiliary datasets. In the transfer of these datasets, its relatedness and importance for the learning of the target model should be considered. Another issue is the heterogeneity of auxiliary datasets. Structure learning of subject-specific causal PGMs could take advantage of auxiliary domains that only contain observational data, or that contain a combination of experimental and observational data. However, since they could come from different subjects, collected conditions may be diverse, such as sampling methods or different sets of intervened variables. Moreover, each dataset encodes specific causal information about the state of a particular subject. The heterogeneity of data should be considered in their transference and fusion. It is of interest for this research to propose a computational solution that addresses this issue.

In summary, this research will address the problem of recovering the structure of subjects-specific causal PGMs $\mathcal{G}_T$ from the knowledge of target subject domain $\mathcal{D}_T$, and transferring knowledge of related source domains $\mathcal{D}_S$ and tasks $\mathcal{T}_S$. Considering that a domain $\mathcal{D} = (\mathbf{X}, D)$ is composed of a set of variables $\mathbf{X}$ and a dataset with samples of $\mathbf{X}$. And a task $\mathcal{T} = (\mathcal{G}, \Theta)$, includes a causal PGM defined over $\mathbf{X}$. More specifically, in this research, assuming that $\mathbf{X}_T = \mathbf{X}_S$, and $\mathcal{T}_T \neq \mathcal{T}_S$, the following conditions will be explored:

- Target domain $\mathcal{D}_T$ with a limited observational dataset: Data from a target subject that were collected without performing manipulations over a $\mathbf{X}_T$.

- Source domains $\mathcal{D}_S$ with observational and experimental datasets: Data from several subjects that were collected after manipulating some subsets of $\mathbf{X}_T$. The set of manipulated variables may vary across the subjects, inclusive may be an empty set.

- Source tasks $\mathcal{T}_S$ with the structure of subject-specific causal PGMs: The structure of causal PGMs for several subjects learned in the past.

1.4 Research Questions

The following questions will guide this doctoral research:

1. Considering that only a limited observable dataset is available for the target subject. How should instances be transferred of auxiliary observational datasets to find the Markov equivalence class, under the assumptions of causal sufficiency, and faithfulness, that best approximate to the true causal structure of a subject-specific PGM?

2. Having learned an approximation to the structure of subject-specific causal PGM with observational datasets, how should causal relations of causal PGMs from auxiliary tasks be transferred to improve this learned causal structure?

3. How should instances of auxiliary datasets with observational and interventional samples be transferred to perform the learning of subject-specific causal PGMs?

4. How should the knowledge transfer of the source domains and tasks be combined to learn subject-specific causal PGMs?

1.5 Hypothesis

Given a limited observable dataset for the target subject, together with auxiliary sources formed by observational and experimental datasets, and causal PGMs:

An algorithm to learn the structure of subject-specific causal PGMs, transferring instances and causal relations from auxiliary sources, under appropriated conditions and considering their variations because of differences in experimental conditions, yields a causal structure for a target subject that better approximates to the true causal structure.

Where *better approximate* means that the obtained causal structure will have a higher number of correctly oriented edges than those obtained by an algorithm that only uses all available data for the target subject.

1.6 Objectives

1.6.1 General Objective

The objective of this research is to develop and validate an algorithm to learn the structure of subject-specific casual PGMs transferring instances from auxiliary datasets with observational and experimental samples and causal relations of auxiliary causal PGMs.

1.6.2 Specific Objectives

1. To develop and validate an instance-based transfer algorithm for learning the Markov equivalence class of the structure of subject-specific causal PGMs using auxiliary observational datasets.

2. To develop and validate a model-based transfer learning algorithm for learning the structure of subject-specific causal PGMs.

3. To develop and validate an instance-based transfer algorithm for learning the structure of subject-specific causal PGMs from auxiliary datasets with a combination of observational and experimental samples.

4. To develop and validate a hybrid algorithm for learning the structure of subject-specific causal PGMs that combine the instances transfer of auxiliar datasets with the model transfer of auxiliar causal PGMs.

5. To exemplify the knowledge transfer learning algorithm, applying it to causal analysis in the neuroscience domain.

1.7 Contributions

- An algorithm that helps to identify the undefined causal directions of Markov equivalent models transferring causal relations of auxiliar causal PGMs under appropriate conditions.

- An instance-based transfer learning algorithm that finds the subject-specific causal PGM that adjusts with the combination of observable instances of target dataset with observable and experimental instances of auxiliar datasets.

- An hybrid knowledge transfer learning algorithm that integrates the instances and model transfer for learning the structure of subject-specific causal PGMs.

1.8 Scope and Limitations

It is considered in this research that a subject is an experimental unit from an experiment of a particular domain. Moreover, it is assumed that all auxiliary datasets and causal PGMs come from the same experiment that was performed under different conditions. Hence, auxiliar datasets, auxiliar causal PGMs, and target dataset share the same set of variables. Besides it is assumed that although there are variations in the causal relations across subjects, auxiliar domains and target subjects share a set of causal relations. Therefore, the algorithm to be developed will recover the structure of target causal PGMs reusing instances and causal relations of auxiliary domains under the interventionist framework proposed by Pearl (2000).

1.9 Structure of the Proposal

The research proposal document has been organized in five chapters. Fundamental concepts for the research are presented in Chapter 2. Several related works are analyzed in Chapter 3. In Chapter 4 the methodology is presented. Finally, some preliminary results are shown in Chapter 5.

Chapter 2

Theoretical Basis

In this chapter, concepts related with Probabilistic Graphical Models and Transfer Learning are presented.

2.1 Probabilistic Graphical Models

Probabilistic graphical models (PGMs) are a compact, efficient and understandable representation of a joint probability distribution of a set of variables. In these graphical models, nodes represent the variables of a domain, and edges, the probabilistic relations between the variables.

Probabilistic graphical models include directed models that upon adding the causal semantics, are used to represent causal relations. The theory related to these PGMs, known as causal Bayesian Networks, will be described in this section.

2.1.1 Probability

Random variables might take a countable (called **discrete variables**) or an uncountable number of possible values (called **continuos variables**). For a set of discrete random variables, their joint probability distribution is defined as follows:

Definition 2.1. (Neapolitan, 2004) The **joint probability distribution of a set of discrete random variables**, $\mathbf{X} = \{X_1, X_2, ..., X_n\}$, is a function that assigns a real value to each combination of the values of the variables in $\mathbf{X}$ and satisfies the following conditions:

1. For each combination of the values of the variables in $\mathbf{X}$,
$$0 \leq P(x_1, x_2, ..., x_n) \leq 1.$$
2. $$\sum_{x_1, x_2, ..., x_n} P(x_1, x_2, ..., x_n) = 1.0.$$

When $\mathbf{X}$ is a set continuous random variables, a density joint function is used to define the joint probability distribution as follows:

Definition 2.2. (Koller *et al.*, 2009) A function $g(\cdot)$, called joint density function, specifies a **joint probability distribution of a set of continuous random variables**, $P(X_1, X_2, ..., X_n)$, if:

1. For all values $x_1, x_2, ..., x_n$,

$$g(x_1, x_2, ..., x_n) \geq 0.$$

2. $g(\cdot)$ is an integrable function, such that,

$$\int_{x_1,x_2,...,x_n} p(x_1, x_2, ..., x_n)dx_1...dx_n = 1$$

3. For any choice of $a_1, a_2, ..., a_n$ and $b_1, b_2, ..., b_n$,

$$P(a_1 \leq X_1 \leq b_1, ..., a_n \leq X_n \leq b_n) = \int_{a_1}^{b_1} ... \int_{a_n}^{b_n} g(x_1, x_2, ..., x_n)dx_1...dx_n$$

The differences between two joint probability distributions could be estimated using the Kullback-Leibler divergence (KLD) in the following form:

Definition 2.3. (Campos, 2006) Given two joint probability distributions defined over $\mathbf{X} = \{X_1, X_2, ..., X_n\}$, $P_E(\mathbf{X}), P_S(\mathbf{X})$, the KLD difference between P_E and P_S is,
$D_{KLD}(P_E, P_S) = \sum_{x_1,x_2,...,x_n} = P_E(x_1, x_2, ..., x_n)log\left(\frac{P_E(x_1,x_2,...,x_n)}{P_S(x_1,x_2,...,x_n)}\right)$

Between the random variables of a probabilistic model, there may be relations of independence which are defined as,

Definition 2.4. (Koller *et al.*, 2009) A variable X is independent of variable Y in a distribution P, written as $X \perp\!\!\!\perp Y$, iff $P(x, y) = P(x)P(y)$ for all values x, y.

Particularly, PGMs represent a set of conditional independence between random variables which are defined as,

Definition 2.5. (Koller *et al.*, 2009) In a distribution P, a variable X is conditionally independent of variable Y given a variable Z, written as $X \perp\!\!\!\perp Y|Z$, iff $P(x, y|z) = P(x|y)P(y|z)$ for all values x, y, z.

2.1.2 Graphs

Others important concepts for PGMs are those related to graph theory.

Definition 2.6. A **graph** is a pair $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ formed by a set of nodes $\mathbf{V} = \{V_1, V_2, ..., V_N\}$, and a set of edges $\mathbf{E} \subset \mathbf{V} \times \mathbf{V}$.

Two nodes are **adjacent** in a graph $\mathcal{G}$, if there is an edge associating them. An edge $(V_i, V_j) \in \mathbf{E}$, may be undirected $(V_i - V_j)$ or directed $(V_i \rightarrow V_j)$ depending whether the order matter, but not of both types. If there is an edge between each pair of nodes in a graph, it is called **complete graph**. When a graph only has undirected edges is called **undirected graph**, and if it only contains directed edges is called a **directed graph**. In a directed edge in the form $V_1 \rightarrow V_2$, V_1 is said to be the **parent** of V_2, and V_2, the **child** of V_1. The set of parents of a node V is denoted as $\mathbf{Pa}(V)$.

Within a graph $\mathcal{G}$, a **path** between two nodes V_0 and V_k is formed by a sequence of nodes, $(V_0, V_1, ..., V_k)$, starting at V_0 and ending at V_k, where $k \geq 1$ and $V_i, V_{i+1} \in \mathbf{E}$ for $i = 0, 1, ..., k-1$. This pair of nodes V_i, V_{i+1} in the sequence is said to be **subsequent nodes**. If there is an undirected edge for each pair of subsequent nodes in a path, it is named **undirected path**; and named **directed path**, if there is a directed edge. A directed path from V_0 to V_k together with the edge $V_k \rightarrow V_0$ form a **directed cycle**. A directed graph in which there are no directed cycles is called a **directed acyclic graph** (DAG). If an acyclic graph contains directed and undirected edges, it is called a **partially directed graph** (PDAG). If there is an undirected edge for each pair of subsequent nodes in a path, it is named **undirected path**; and named **directed path**, if there is a directed edge. A directed path from V_1 to V_2 together with the edge $V_2 \rightarrow V_1$ form a **directed cycle**. A directed graph in which there are no directed cycles is called a **directed acyclic graph** (DAG). If an acyclic graph contains directed and undirected edges, it is called a **partially directed graph** (PDAG).

The undirected graph resulting from ignoring the direction of edges in a DAG is the **skeleton** of the DAG. A **v-structure** in a DAG is an ordered triple of nodes (X, Y, Z), such that, the edges $X \rightarrow Y$ and $Y \leftarrow Z$ are in the DAG, and there is no edge between the nodes X, Z (Chickering, 2002). Two DAGs are equivalent if and only if they have the same skeletons and the same v-structures (Kalisch & Bühlmann, 2014). A set of equivalent directed acyclic graphs is called a **Markov equivalence class** (He *et al.*, 2015).

Within a DAG, it is possible to identify conditional independences between random variables, using a criterion known as **d-separation**,

Definition 2.7. (Spirtes *et al.*, 2000) In a DAG $\mathcal{G}$, if X and Y are nodes, $X \neq Y$, and $\mathbf{W}$ a set of nodes that does not contain X or Y, then X and Y are **d-separated** given $\mathbf{W}$ in $\mathcal{G}$, iff there exists no undirected path U between X and Y[1], such that, every node V on U in the form $V_1 \rightarrow V \leftarrow V_2$, has a descendent in $\mathbf{W}$, and no other node on U is in $\mathbf{W}$.

When $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{W}$ are disjoint sets of nodes in $\mathcal{G}$, and $\mathbf{U} \neq \emptyset, \mathbf{V} \neq \emptyset$, it said that $\mathbf{U}$ and $\mathbf{V}$ are d-separated iff every pair $(U, V) \in \mathbf{U} \times \mathbf{V}$ is d-separated given $\mathbf{W}$.

2.1.3 Associative Probabilistic Graphical Models

Definition 2.8. (Sucar, 2015) A probabilistic graphical model defined over a set of variables $\mathbf{V} = \{X_1, X_2, ..., Xn\}$, is a pair $(\mathcal{G}, \Theta)$, where $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ is the graph, with $\mathbf{E} \subset \mathbf{V} \times \mathbf{V}$ that represents the structure of the model; and $\Theta = \{\theta(\mathbf{y}_i)\}, \mathbf{Y}_i \subset \mathbf{V}$ is the set of local functions that defines the parameters of the model. The product of local functions defines the joint probability distribution of $\mathbf{V}$.

Bayesian Networks are a type of Probabilistic Graphical Models in which their structure is formed by DAGs, formally:

[1] A resulting path between X and Y after ignoring the direction of edges.

Definition 2.9. (Koller *et al.*, 2009) A **Bayesian Network** (BN) defined over a set of variables $\mathbf{V} = \{X_1, X_2, ..., X_n\}$ is a pair $(\mathcal{G}, \Theta)$ where $\mathcal{G}$ is a directed acyclic graph, and $\Theta = \{\theta_i\}$ is the set of local functions in the form $\theta_i = P(x_i|\mathbf{pa}(X_i))$.

Parametrically, a Bayesian network represents the joint probability distribution over $\mathbf{V}$. The structure of Bayesian networks represents a set of conditional independences. Each node in a BN is conditionally independent of its non-descendants given its parents. This last property allows to estimate the joint probability distribution over $\mathbf{V}$ in the form:

$$P(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} P(x_i|\mathbf{pa}(X_i)) \tag{2.1}$$

This factorization of the P($\mathbf{V}$) relative to a graph $\mathcal{G}$ is known as Markov compatibility. Formally,

Definition 2.10. (Pearl, 2000) If a probability function P admits the factorization given by equation 2.1, relative to a DAG $\mathcal{G}$, is said that P represents $\mathcal{G}$, and, P is **compatible** or **Markov relative** to $\mathcal{G}$.

2.1.4 Causal Probabilistic Graphical Models

According to Spirtes *et al.* (2000), **causation** is a relation between two particular events where there is an event that generates an effect on another one. It is defined that causation is a relation:

Transitive: If an event A causes and event B, and then B causes another event C, then A causes C.

Irreflexive: an event cannot cause itself.

Antisymmetric: if an event A causes an event B, then B is not a cause of A.

Information about causal relations between variables is encoded in causal probabilistic graphical models[2]. In particular, Bayesian Networks are used for modeling causal relations, because they provide facilities to represent and infer effects of actions. In order to Bayesian networks encode reliably causal relations, a set of assumptions is required in their construction that is summarized in the following definition:

Definition 2.11. (Pearl, 2000) Let $P(\mathbf{v})$ be a probability distribution over a set $\mathbf{V}$ of variables, and let a $P(\mathbf{v}|do(\mathbf{X} = \mathbf{x}))$ denote the resulting distribution from the intervention $do(\mathbf{X} = \mathbf{x})$ that sets a subset $\mathbf{X}$ of variables to constants $\mathbf{x}$, and delete all incoming edges to $\mathbf{X}$. Denote by $\mathbf{P}*$ the set of all interventional distributions $P(\mathbf{v}|do(\mathbf{X} = \mathbf{x}))$, $\mathbf{X} \subseteq \mathbf{V}$, including $P(\mathbf{v})$ that represents no intervention (i.e. $\mathbf{X} = \emptyset$). A DAG $\mathcal{G}$ is said to be a **causal Bayesian network** (CBN) compatible with $\mathbf{P}*$ if and only if the following three conditions hold for every distribution in $\mathbf{P}*$:

1. $P(\mathbf{v}|do(\mathbf{X} = \mathbf{x}))$ is Markov relative to $\mathcal{G}$;

[2] Although there are several types of causal models, in this research we consider the causal PGMs.

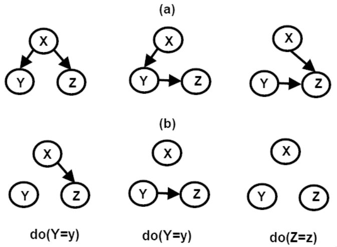


Figure 2.1: An example of (a) basic causal graphs and (b) its corresponding modified causal graph due to the intervention.

2. $P(v_i|do(\mathbf{X}=\mathbf{x})) = 1 \ \forall V_i \in \mathbf{X}$ whenever v_i is consistent with $\mathbf{x}$;

3. $P(v_i|do(\mathbf{X}=\mathbf{x}), \mathbf{pa}(V_i)) = P(v_i|\mathbf{pa}(V_i)) \ \forall V_i \notin \mathbf{X}$ whenever $\mathbf{pa}(V_i)$ is consistent with $\mathbf{x}$.

The set of assumptions considered in this definition allow differentiating Bayesian networks from causal Bayesian networks. This definition assumes that the structure $\mathcal{G}$ of a causal Bayesian Network complies with the following rule:

Definition 2.12. (Spirtes *et al.*, 2000) $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ represents a **causal graph**, when there is a directed edge $X \to Y$ in $\mathbf{E}$ iff X is a direct cause of Y relative to $\mathbf{V}$.

Considering that a direct cause between variables is defined as follows:

Definition 2.13. (Zhang & Spirtes, 2008) X is a **direct cause** of Y relative to $\mathbf{V}$, if there exists $x_1 \neq x_2$ and $\mathbf{z}$ with $\mathbf{Z} = \mathbf{V} \setminus \{X, Y\}$, such that $P(y|do(X = x_1), \mathbf{Z} = \mathbf{z}) \neq P(y|do(X = x_2), \mathbf{Z} = \mathbf{z})$.

In this definition of direct cause, an intervention $do(X = x)$ is considered a mechanism that fixes X to value x, and delete all direct causes over X (the incoming edges to X). The reason for this graph surgery is due to these direct causes of X have no influence during the intervention. Moreover, the intervention makes the intervened variable independent of its direct causes. In Figure 2.1 are shown some examples of interventions.

Finally, in the definition of causal Bayesian networks is assumed that is possible to estimate the effects of interventions from the pre-intervened joint probability distribution $P(\mathbf{v})$ and the graph of the causal Bayesian network. Considering that an intervention $do(X_i = x'_i)$ transforms the causal graph and the pre-intervened joint probability distribution $P(\mathbf{v})$, the interventional probability distribution is estimated as follows:

$$P(\mathbf{v}|do(X_i = x'_i)) = \begin{cases} \prod_{j \neq i} p(x_j|\mathbf{pa}(X_j)), \text{if } x_i = x'_i \\ 0, \text{if } x_i \neq x'_i \end{cases} \tag{2.2}$$

Where $P(\mathbf{v}) = \prod_{x_i} P(x_i|\mathbf{pa}(X_i))$ is obtained from passive observations of the set $\mathbf{V}$ of the causal system (without performing any intervention).

In summary, causal Bayesian networks are PGMs more expressive than Bayesian networks, since in addition to encode probabilistic dependencies between the variables, they encode causal information. An example of a causal BN with the statistical and causal relations represented in it is presented in Figure 2.2.

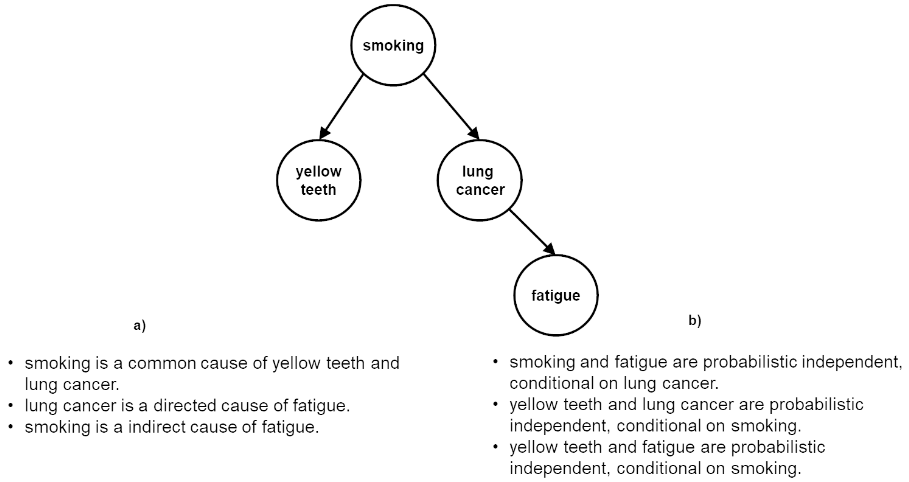

smoking
yellow teeth
lung cancer
fatigue
a)
b)
• smoking is a common cause of yellow teeth and lung cancer.
• lung cancer is a directed cause of fatigue.
• smoking is a indirect cause of fatigue.
• smoking and fatigue are probabilistic independent, conditional on lung cancer.
• yellow teeth and lung cancer are probabilistic independent, conditional on smoking.
• yellow teeth and fatigue are probabilistic independent, conditional on smoking.

Figure 2.2: Example of causal Bayesian Network with some of (a) the causal and (b) the probabilistics relations represented.

With a causal Bayesian network it is possible to formulate three types of questions (Pearl & Mackenzie, 2018):

Associative: This type of question involves seeing and observing the environment. An example query is, what does a symptom tells me about a disease?

Interventional: This type of question involves doing and intervening in the environment. These are prospective questions about what are the effects if we intervened the environment. An example query is, what if I take aspirin, will my headache be cured?

Counterfactual: This type of question involves the actions of imagining, retrospection, understanding the environment. These are retrospective questions about what would have happen if we took another action than one we are currently observing. An example query is, what if I had not been smoking?

2.1.5 Structure Learning of Causal Bayesian Networks

The learning of causal Bayesian networks is often performed in two steps: learning of the causal structure, and estimation of parameters from data and the causal structure. The type of datasets used in the learning of the causal structure may be:

Observational: Data corresponding to measurements made under natural conditions of a causal system.

Experimental: Data corresponding to measurements made under different disturbances of the system caused by external interventions.

Ideally, experimental data should be used in the structure learning of causal BNs. However, these data are difficult to collect, because it is complex, costly, or time-demanding to

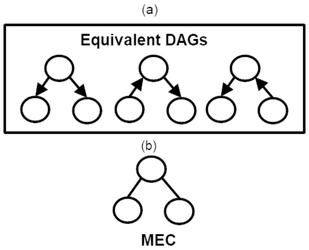


Figure 2.3: An example with (a) a set of equivalent DAGs and (b) its corresponding MEC.

perform experiments. Therefore, several algorithms have been developed to learn partial causal structures only from observational data. These algorithms, known as causal discovery algorithms, often relies on a number of assumptions including but not limited to the following (Spirtes & Zhang, 2016):

Causal Sufficiency (CS) Every common cause of two or more variables in a set of variables $\mathbf{V}$, also is in $\mathbf{V}$.

Causal Markov Condition (CMC) Each variable in the causal structure is independent of its non-effects given its directed causes.

Faithfulness Condition (FC) Each true conditional independence between variables is entailed by the causal structure.

Causal Minimality Condition (CLC) No proper subgraph of the true causal $\mathcal{G}$ over $\mathbf{V}$, with joint distribution P, satisfies that P is Markov relative to $\mathcal{G}$.

The causal discovery algorithms, from the analysis of observational data and under several assumptions, can only recover a set of structures. Specifically, directed acyclic graphs equivalents to the true structure of a causal BN are recovered and grouped in a Markov equivalence class (MEC). An example of a set of Markov equivalent DAGs and the corresponding MEC is shown in Figure 2.3.

Three types of discovery causal algorithms have been proposed: Constraint-based, Score-based, and based on functional causal models (Glymour *et al.*, 2019; Zhang *et al.*, 2018).

Constraint-based algorithms

These algorithms perform hypothesis tests to search the MEC that is consistent with the conditional independence found in the data. Popular algorithms of this type include PC (Spirtes & Glymour, 1991) and its variant, the Fast Causal Inference algorithm (FCI) (Spirtes *et al.*, 1995). PC recovers a MEC that includes the true causal structure under CS, CMC and FC assumptions. While FCI does not require CS and recovers MECs under CMC and FC assumptions. These algorithms use statistical tests of conditional independence, which require a representative sample size to be reliable (Zhang & Spirtes, 2016). In addition, some typical tests impose restrictions over the distribution of the data (Malinsky & Danks, 2018). PC starts with a complete graph, and in each iteration, the algorithm deletes edges when the pairs of variables are conditionally independent given a subset of variables. In each iteration of the algorithm, the size of conditional variables subset is incremented, until there are no pair of adjacent nodes, (X, Y), in which, all variables in the conditional subsets are adjacents to X or Y (Glymour *et al.*, 2019).

Score-based algorithms

These algorithms find the MEC that includes the true causal graph by optimizing a score function. In the space of MECs, the MEC with the highest score is searched. Algorithms, such as Greedy Equivalence Search (GES) (Chickering, 2002) including its extensions Greedy Fast Causal Inference (GFCI)(Ogarrio *et al.*, 2016), Fast GES (FGES) (Ramsey *et al.*, 2017) perform a local heuristic search that reliably finds the best MEC in the limit of infinite data and under the four causal assumptions (Guo *et al.*, 2018; Zhang *et al.*, 2018), with the exception of GFCI that does not require CS.

GES is a two-stage algorithm that starts with an empty graph and heuristically searches for adding and deleting edges that improve the score function. Best edges are added in the first stage, and in the second stage, edges which improve the score function are removed. The stages of GES algorithm are described in Algorithm 1.

In the GES algorithm, Bayesian Dirichlet equivalent and Uniform (BDeU) score function is used for learning MECs defined over discrete variables. BDeU score estimates the probability that a dataset D adjust with the structure $\mathcal{G}$ of a PGM as follows (Chickering, 2002):

$$BDeU(\mathcal{G}, D) = score(\mathcal{G}, D) = \prod_{i=1}^{n} \left\{ \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})} \right\} \tag{2.3}$$

where,

$\Gamma(\cdot)$ is the Gamma function,

n is the number of nodes in $\mathcal{G}$,

q_i is the number of values of $\mathbf{Pa}_T(X_i)$,

r_i is the number of values of X_i,

N_{ijk} is the number of cases in which $X_i = k$ and its parents $\mathbf{pa}(X_i) = j$,

$N_{ij} = \sum_k N_{ijk}$,

$\alpha_{ijk} = \frac{\alpha}{r_i q_i}$ is a Dirichlet prior parameter, and $\alpha_{ij} = \sum_k \alpha ijk$.

BDeU score assigns the same score to equivalent structures and also it is descomposable at node level. It can be expressed as product of local scores in the form:

$$localScore(X_i, \mathbf{Pa}(X_i), D) = \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})} \tag{2.4}$$

This local score estimates the adjustment of the data D with a local structure formed by $X_i \in \mathbf{X}$ and its parents $\mathbf{Pa}(X_i)$.

Algorithm 1: GES algorithm

```
Algorithm GES()
    Input: $\mathbf{V}$
    $D$: a observational dataset from $\mathbf{V}$

    Output: $\mathcal{G} = (\mathbf{V}, \mathbf{E})$: A partially
            directed graph
    $\mathcal{G} \leftarrow \emptyset$
    $s_{best} \leftarrow -\infty$
    /* First stage:  Adding edges    */
    repeat
        foreach $E' = (X, Y), E' \notin \mathbf{E}$ do
            if $validInsert(E', \mathcal{G})$ then
                $\mathcal{G}' \leftarrow (\mathbf{V}, \mathbf{E} \cup \{E'\})$
                $s \leftarrow score(\mathcal{G}', D)$
                if $s > s_{best}$ then
                    $s_{best} \leftarrow s$
                    $E_{best} \leftarrow E'$
                end
            end
        end
        $\mathbf{E} \leftarrow \mathbf{E} \cup \{E_{best}\}$
    until $(s_{best} < s_{ant})$
    /* Second stage:  Deleting edges */
    repeat
        $s_{ant} \leftarrow s_{best}$
        foreach $E' = (X, Y), E' \in \mathbf{E}$ do
            if $validDelet(E', \mathcal{G})$ then
                $\mathcal{G}' \leftarrow (\mathbf{V}, \mathbf{E} \setminus \{E'\})$
                $s \leftarrow score(\mathcal{G}'(\mathbf{V}, \mathbf{E}'), D)$
                if $s > s_{best}$ then
                    $s_{best} \leftarrow s$
                    $E_{best} \leftarrow E'$
                end
            end
        end
        $\mathbf{E} \leftarrow \mathbf{E} \setminus \{E_{best}\}$
    until $(s_{best} < s_{ant})$
    return $\mathcal{G}$

Function validInsert($(X, Y), \mathcal{G}$)
    $\mathbf{T} \leftarrow \{T\}, (T, Y) \in \mathcal{G}, (T, X) \notin \mathcal{G}$
    $\mathbf{NAYX} \leftarrow \{Z\}, (Z, Y), (Z, X) \in \mathcal{G}$
    if $\mathbf{T} \cup \mathbf{NAYX}$ is a complete subgraph and
     every semidirected path from $Y$ to $X$
     does not contain any node in $\mathbf{NAYX}$
     then
        return True
    end
    return False

Function validDelet($(X, Y), \mathcal{G}$)
    $\mathbf{NAYX} \leftarrow \{Z\}, (Z, Y), (Z, X) \in \mathcal{G}$
    $\mathbf{H0} \leftarrow \mathbf{NAYX}$
    foreach $\mathbf{H} \subset \mathbf{H0}$ do
        if $\mathbf{NAYX} \setminus \mathbf{H}$ not is a complete
         subgraph then
            return False
        end
    end
    return True
```

Algorithms based on functional causal models

In these algorithms causal relations are described by functional models in the form of $Y = f(X, \epsilon)$, where ϵ is the noise term, and X, Y are the variables cause and effect, respectively (Zhang *et al.*, 2018). Linear models (Montero-Hernandez *et al.*, 2018), Linear Non-Gaussian Model (LiNGaM) (Shimizu *et al.*, 2006), NonLinear Additive Noise Model (ANM) (Hoyer *et al.*, 2009), and the Post-Nonlinear (PNL) causal model (Zhang & Hyvärinen, 2009) are typical functional models. Algorithms based on functional causal models learn causal relations with more detail, and in some cases, when some appropriate constraints are imposed to the functional model, a unique DAG within a Markov equivalence class can be identified (Zhang *et al.*, 2018).

2.2 Transfer Learning

Transfer learning is a methodology to learn models from data of different domains. Its main aim is to improve the performance of models by leveraging knowledge from auxiliary domains. Consequently, it is a desirable methodology to solve problems in domains where collecting data is difficult or impossible.

Domains and tasks are two important elements in the transfer learning methods. A **domain**, $\mathcal{D} = (\mathcal{X}, P(\mathbf{X}))$, is composed by two parts: a feature space $\mathcal{X}$ and a marginal probability distribution $P(\mathbf{X})$, where $\mathbf{X} = \{X_1, X_2, ..., X_n\}$ is a set of features. A **task**, $\mathcal{T} = (\mathcal{Y}, f(\cdot))$ is formed by a label space $\mathcal{Y}$, and a predictive function $f(\cdot)$ which could be learned from a training dataset $D = \{(x_i, y_i)\}$ composed by pairs $(\mathbf{x}_i, y_i), \mathbf{x}_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$ (Weiss *et al.*, 2016).

Considering the definitions of domain and task, a formal definition of transfer learning is as follows:

Definition 2.14. (Pan & Yang, 2010) Given a source domain $\mathcal{D}_S$, a source task $\mathcal{T}_\mathcal{S}$, a target domain $\mathcal{D}_T$, and a target task $\mathcal{T}_\mathcal{T}$, **transfer learning** is a process that helps to improve the performance of the predictive function[3] $f(\cdot)_T$ in $\mathcal{T}_\mathcal{T}$, using the knowledge in $\mathcal{D}_\mathcal{S}$ and $\mathcal{T}_\mathcal{S}$, where $\mathcal{D}_\mathcal{S} \neq \mathcal{D}_\mathcal{T}$ or $\mathcal{T}_\mathcal{S} \neq \mathcal{T}_\mathcal{T}$.

The conditions $\mathcal{D}_\mathcal{S} \neq \mathcal{D}_\mathcal{T}$ or $\mathcal{T}_\mathcal{S} \neq \mathcal{T}_\mathcal{T}$ of the transfer learning definition implies the following cases (Aggarwal, 2014):

Case 1) The source and target domains have the same feature space, $\mathcal{X}_\mathcal{S} = \mathcal{X}_\mathcal{T}$, but differ in the marginal distributions, $P(\mathbf{X}_S) \neq P(\mathbf{X}_T)$. Transfer learning under these conditions is called **homogeneous transfer learning**.

Case 2) The source and target domains differ in their feature spaces, i.e. $\mathcal{X}_\mathcal{S} \neq \mathcal{X}_\mathcal{T}$. Transfer learning under these conditions is called **transfer learning across heterogeneous feature spaces**.

Case 3) Target and source tasks differ in their label spaces, i.e. $\mathcal{T}_\mathcal{S} \neq \mathcal{T}_\mathcal{T}$. This is another case of heterogeneous transfer learning known as **transfer learning across heterogeneous label spaces**.

[3] The accuracy of the predictive function.

2.2.1 Issues on Transfer Learning

Ideally, a transfer method must improve the performance of a target task, avoiding transfer knowledge which affects it. This last phenomenon is known as **negative transfer** and formally is defined as follows:

Definition 2.15. (Weiss *et al.*, 2016) Let a source domain $\mathcal{D}_S$, a source task $\mathcal{T}_S$, a target domain $\mathcal{D}_T$, a target task $\mathcal{T}_T$, a predictive function $f_{T1}(\cdot)$ learned with only $\mathcal{D}_T$, and a predictive function $f_{T2}(\cdot)$ learned with a transfer learning process, combining $\mathcal{D}_T$ and $\mathcal{D}_S$. **Negative transfer** occurs when the performance of $f_{T1}(\cdot)$ is greater than the performance of $f_{T2}(\cdot)$.

In addition to the above, the design of effective transfer methods should consider the following three issues:

What to transfer?: It refers to defining which parts of the knowledge (features, parameters of $f(\cdot)$, or labels) could be transferred across domains or tasks. The common knowledge between different domains which helps to improve the learning of a target task should be selected.

How to transfer?: It refers to developing algorithms that combines the knowledge of related domains in the learning of a target task.

When to transfer?: It refers to defining heuristics that identify in which situations the knowledge should and should not be transferred.

2.2.2 Transfer Learning Categories

Transfer learning approaches according to the form of information transfer are classified in the following four categories (Weiss *et al.*, 2016; Aggarwal, 2014):

Instance-based transfer learning: The methods in this category transfer the instances of source domains to a target domain. The instances, after re-weighting or re-sampling in the target domain, are used to learn a target task. Two issues are considered: no labeled data are available, and few data are available.

Model parameter-based transfer learning: In this category, the methods assume that the source and target tasks share some parameters or prior distributions of the hyper-parameters of the models. Hence, some methods reuse the parameters of source tasks in the target predictive function. Other methods, learn multiple models from source domains, and then they are weighted and combined to construct a target model.

Feature-based transfer learning: The aim of these methods is to learn the best feature representation for the source and target domains. This category includes two types of methods: asymmetric feature transformation and symmetric feature transformation. The first category of methods is applied when there are differences between the conditional distributions of source and target domains caused by the difference between the features of source and target domains. The features of source domains are weighted by these methods to reduce these difference. In the second category of methods, finding a set of latent features reduces the difference between the marginal distributions of source and target domains.

Relational-based transfer learning: These methods use a relationship between the source and target domains to construct a target predictive function. The methods in this category assume that some relationships between objects or instances are similar across domains or tasks. Hence common relationships are extracted and re-used in the target task.

2.2.3 Transfer Learning in Probabilistic Graphical Models

In transfer learning for probabilistic graphical models, a domain $\mathcal{D} = (\mathbf{V}, D)$ is composed by a set of variables $\mathbf{V}$ and a dataset D that includes a set of samples from $\mathbf{V}$. A task $\mathcal{T} = (\mathcal{G}, \Theta))$ includes a PGM defined over $\mathbf{V}$. Considering these elements, the problem of transfer knowledge for learning probabilistic graphical models is defined as follows:

Definition 2.16. (Jia *et al.*, 2018; Zhou *et al.*, 2016) Given a source domain $\mathcal{D}_{\mathcal{S}}$, with its corresponding source task $\mathcal{T}_{\mathcal{S}}$, a target domain $\mathcal{D}_T$ with its corresponding target task $\mathcal{T}_{\mathcal{T}}$, **transfer learning** is a process that helps to improve the quality of the probabilistic graphical model in $\mathcal{T}_{\mathcal{T}}$, using the knowledge in $\mathcal{D}_{\mathcal{S}}$ and $\mathcal{T}_{\mathcal{S}}$, where $\mathcal{T}_{\mathcal{S}} \neq \mathcal{T}_{\mathcal{T}}$.

Where the quality of a PGM is the accuracy of its associate joint probability distribution $P(\mathbf{x})$, and the number of correctly oriented edges of the $\mathcal{G}$.

In transfer learning for PGMs, the relation between $\mathcal{D}_{\mathcal{S}}$ and $\mathcal{D}_{\mathcal{T}}$ implies the following cases:

Homogeneous transfer learning: The source and target domains have the same set of variables, $\mathbf{V_S} = \mathbf{V_T}$, but differ in their joint probability distributions, $P(\mathbf{V}_S) \neq P(\mathbf{V}_T)$ (which are estimated from D_S and D_T, respectively) (Jia *et al.*, 2018).

Heterogeneous transfer learning: The source and target domains differ in their sets of variables, i.e. $\mathbf{V_S} \neq \mathbf{V_T}$ (Zhou *et al.*, 2016).

Chapter 3

Related Work

In this chapter, a review of related works with this research are presented. Specifically, algorithms for learning probabilistic graphical models, including transfer learning, learning from multiple datasets, and subject-specific learning, are analyzed.

The concepts of population-wide and subject-specific models are used in the review. Population-wide model is used to refer to models that adjust with the characteristics of the all population, while the subject-specific model, to refer to models that adjust with the special characteristics of a subject of population. In the case of a population-wide causal probabilistic graphical model (in short population-wide causal model), it refers to models encoding the common causal relations and their corresponding local probabilistic distributions to a population. While subject-specific causal probabilistic graphical model (in short subject-specific causal model) is used to refer to models that encode the corresponding causal relations with their probabilistic parameters of a particular individual.

3.1 Transfer for Learning Probabilistic Graphical Models

Learning associative probabilistic graphical models applying transfer learning techniques has been explored by diverse works. For example, Zhou *et al.* (2016) provides a transfer framework to learn the parameters of Bayesian networks. This framework includes a relevance metric and a fusion function that combines optimally the knowledge of relevant domains. Additionally, Luis *et al.* (2010) and Cameras *et al.* (2013) proposed transfer algorithms to learn Bayesian networks and dynamic Bayesian networks, respectively. In these works, a modification of the PC algorithm, using local and global similarity measures for determining the relevance of auxiliary domains in the independence tests between variables, is proposed to learn the structure of BNs. The local parameters are estimated from auxiliary domains which have the same local structure than the target. For its part, Niculescu-Mizil & Caruana (2007), Oyen & Lane (2013), and Oates *et al.* (2016) provide transfer algorithms to learn simultaneously the structure of multiple Bayesian networks. Assuming that all tasks share the same order in the variables, the best structures can be searched using a global Bayesian score (Niculescu-Mizil & Caruana, 2007) or a local Bayesian score (Oyen & Lane, 2013). This restriction is not required for the algorithm of Oates *et al.* (2016) that applies an integer linear programming method to estimate the multiple structures.

The mentioned algorithms were designed to learn associative models that do not consider the framework to learn causal models from observational data: Causal Markov Condition, Faithfulness, Minimality, and Causal Sufficiency conditions. Hence, the PGMs learned with these algorithms cannot be interpreted as causal models. The most related algorithm to our proposal was provided by Jia *et al.* (2018). They propose a homogeneous transfer algorithm for learning the structure of causal Bayesian networks, which assumes Causal Sufficiency, Causal Markov Condition, and Faithfulness conditions. This algorithm, which is a modified PC algorithm, is limited for partially learning target causal structures from auxiliary datasets that have the same relevance and only contain observational data. Their proposed modifications to the PC algorithm are focused on using heuristics for deciding whether auxiliary domain data can be used in conditional independence tests. The learned causal models with this algorithm showed slightly smaller error (for adding, deleting and reversing edges) than those obtained with an algorithm that only concatenates data.

3.2 Learning Causal Probabilistic Graphical Models from Multiple Datasets

Learning of causal probabilistic graphical models from multiple datasets have been explored in several works (Claassen & Heskes, 2010; Hauser & Bühlmann, 2012; Mooij *et al.*, 2019; Ramsey *et al.*, 2010; Tillman & Spirtes, 2011; Triantafillou & Tsamardinos, 2015). The aim of these algorithms is to learn the structure of population-wide causal models from the combination of multiple datasets that were obtained under different conditions. In these works, it is assumed that there exists a single underlying causal mechanism. Moreover, in the large sample limit, their algorithms find the Markov equivalence class that contains the true causal structure.

Algorithms for partially learning causal structures from multiples datasets with observational data were proposed in (Claassen & Heskes, 2010; Ramsey *et al.*, 2010; Tillman & Spirtes, 2011). Ramsey *et al.* (2010) provide a modification of the GES algorithm, where a modified Bayesian score combines the scores that were estimated in each dataset and then find the causal structure. In the proposal of (Claassen & Heskes, 2010; Tillman & Spirtes, 2011), the causal structure for each dataset is first searched, and from these, a summarized causal structure is found. A special case, where there are datasets with observations for some subset of variables of the systems, is studied by Claassen & Heskes (2010); Tillman & Spirtes (2011). They assume that the datasets have variables in common, and at least one dataset measures every variable of the system.

On another hand, Hauser & Bühlmann (2012); Mooij *et al.* (2019); Triantafillou & Tsamardinos (2015) analyzed causal structure learning from a combination of experimental and observational datasets. Hauser & Bühlmann (2012); Mooij *et al.* (2019), use context variable to indicate the intervened variables and from a single dataset, including context and observable variables, learn the causal structure. Their proposals assume causal sufficiency. In special, the algorithm of Hauser & Bühlmann (2012), that is a modification of the GES algorithm, called Greedy Interventional Equivalence Search (GIES), also requires knowing what variables were manipulated in the experiment. While the framework proposed by Mooij *et al.* (2019) is designed to work with constrained-based algorithms and includes a strategy to construct a single dataset from multiples datasets with different subsets of intervened variables. For its part, Triantafillou & Tsamardinos (2015) provides a constrained-based algorithm that can find the causal structure from datasets including samples for a subset of the variables, with common variables between them. A causal structure from each dataset is estimated applying the FCI algorithm, and then a summarized causal structure that fits with all datasets is found. For finding the summarized causal structure, the set of independences and dependences entailed for each dataset, and a strategy to manipulate conflicts are used.

The mentioned algorithms were designed to learn causal PGMs under the assumption that there is a sufficient number of samples (Zuk et al., 2012). In case of limited dataset, these algorithms can not learn acceptable causal PGMs.

3.3 Learning Subject-Specific Probabilistic Graphical Models

The issue of learning subject-specific probabilistic graphical models was considered by Visweswaran & Cooper (2010), Cooper *et al.* (2018), Jabbari *et al.* (2018), and Li *et al.* (2018). In (Visweswaran & Cooper, 2010) an algorithm to learn patient-specific associative Markov Blanket models is proposed. In this algorithm, Markov Blanket models are estimated by averaging a set of selected Bayesian networks. Using the knowledge of a subject-specific instance and a single training dataset, a candidate set of BNs are learned. A score based on Kullback–Leibler (KL) divergence estimates the relevance of a candidate BN for finding the best Markov Blanket model.

For its part, Cooper *et al.* (2018), Jabbari *et al.* (2018), and Li *et al.* (2018) consider the learning of subject-specific causal models. The algorithm of Cooper *et al.* (2018) learns specific causal models represented by bipartite causal graphs, from observational datasets, assuming that there is only a cause for each effect.

Li *et al.* (2018) provides an algorithm to find the structure of a subject-specific causal model from observable and interventional datasets. This algorithm assumes that the causal effects vary across subjects while the direction of causal relations is homogeneous. For modeling the subject-specific causal models, linear Gaussian functional equations with mixed effects are used. The subjects-specific functional models are learned from observational and interventional datasets of the same subject.

On the other hand, Jabbari *et al.* (2018) provided the first algorithm to learn the structure of instance-specific causal Bayesian networks that is called instance-specific GES (IGES). They propose an algorithm that is limited for learning partial causal PGMs from one observational dataset and an instance that describes special characteristics of a specif subject. It is a context-based algorithm in which it is considered that certain independences are holding in a specific assignment of values for specific variables. Under this assumption, besides those of Causal Markov condition, Faithfulness and Causal Sufficiency, a partial causal model that best adjust with the characteristics of a specific-subject represented by an instance is found. First, the algorithm performs a greedy equivalence search (GES) with a single observational dataset for finding the population-wide causal Bayesian networks. Then, using the population-wide causal model as *a priori* model and a subject-specific instance, the instance-specific partial causal model (a Markov equivalence class) is found with a modified version of the GES algorithm. It is considered that the subject-specific instance is not included in the training dataset. Hence, the algorithm, using the training set, searches the independencies that are consistent with the context described by the subject-specific instance. With a high adjacency (0.7 ± 0.07), and a low arrowhead (0.37 ± 0.13) average precision, the algorithm finds the specific causal relations for a particular subject.

3.4 Discussion

Although there are several works that have explored knowledge transfer for learning Probabilistic Graphical Models, the research for causal PGMs remains limited. Most studies have relied on the learning of associative PGMs. To our knowledge, only the work of Jia *et al.* (2018) have examined the transfer learning of causal PGMs. The transfer algorithm provided by Jia *et al.* (2018) is limited to work with auxiliary observational datasets that have the same relevance for learning a target causal PGM. Moreover, this algorithm assumes that the variations in the datasets are due to sampling, ignoring variations due to differences in experimental conditions.

A similar situation occurs with works for learning subject-specific causal models. The algorithm of Jabbari *et al.* (2018) is constrained to find the model that best adjusts with the characteristics described by an instance of the target, using one observational dataset. The possible differences of the target subject with members of the population are not contemplated in the learning. Moreover, since this algorithm use observational data, it is limited to find partial causal models.

On another hand, although there are numerous studies that have analyzed the learning of causal models from multiple datasets, they are focused on the learning of population-wide causal models. Besides, since these algorithms assume that there are enough data for learning causal PGMs, they cannot be used for learning acceptable causal PGMs from limited datasets.

Chapter 4

Methodology

4.1 Working Plan

The following methodology is proposed to achieve the objectives of this research:

1. **Assessment of current solutions for structure learning of causal PGMs**.

   In this step it will be studied the feasibility of extending causal structure discovery algorithms with transfer learning techniques. In particular, the score-based algorithms GES, GIES, GFCI, and IGES will be analyzed. Properties of score functions, causal assumptions, and type of Markov equivalence class will first be identified. Then, a preliminary knowledge transfer algorithm for learning the structure of subject-specific causal PGMs from two auxiliary observational datasets and a one auxiliary causal PGM will be designed.

2. **Design and development of instance-based transfer algorithm for learning Markov equivalence classes**.

   In this stage, a score-based algorithm for learning MECs, transferring instances from auxiliar observational datasets will be developed. The aim is to develop an algorithm for finding a preliminary causal structure encoding the possible causal relations for a subject-specific causal PGM.

   a) Design a relevance function that measure the relation of the auxiliary domain with the target domain, and how much the auxiliary dataset helps to learn a target causal PGM. This function will be defined over the differences in local conditional distributions.

   b) Design a scoring function (Chickering, 2002) that measure how well the local structure of target causal PGM fits with the combination of auxiliary and target datasets.

   c) Design an algorithm to combine auxiliary domains considering the relevance of each auxiliary datasets. In its design, two scenarios will be analyzed, transferring weighted instances of the best relevant domains, and combining the causal models learned independently from each auxiliary dataset.

d) Validation. Its ability for recovering skeleton and v-structures of the ground truth causal structures will be assessed. To do this evaluation, different causal PGMs with known structure and parameters will be used as ground truth causal PGMs. Next, from these ground truth causal PGMs, auxiliar causal PGMs will be generated. Target datasets and auxiliary datasets will be sampled from the ground truth causal PGMs and auxiliar causal PGMs, respectively. Finally, the preliminary causal subject-specific PGMs (MECs) obtained by the algorithm will be compared with the ground truth causal PGMs. This evaluation process is illustrated in Figure 4.2.

i) Generation of synthetic datasets: Following the scheme proposed in Luis *et al.* (2010), datasets for target and auxiliar domains will be generated. In this scheme, the observational dataset for target subjects are sampled from the ground truth causal PGMs, and the auxiliary datasets are sampled from auxiliar causal PGMs. Adding and deleting edges of a ground truth causal PGM, the auxiliar causal PGMs are generated.

ii) An experiment that contemplates different ground truth causal PGMs and different modification schemes for generating auxiliar datasets will be designed and performed. For evaluating the MECs obtained by the algorithm, normalized structural Hamming distance (NSHD), and adjacency precision (Jabbari *et al.*, 2018; Triantafillou & Tsamardinos, 2015), will be used.

3. **Design and development of model-based transfer learning algorithm**.

In this stage, an algorithm that helps to identify the direction of the causal relations for preliminary subject-specific causal PGMs (MECs), will be developed and validated. The algorithm will be designed considering the theory and conditions defined in (Bareinboim & Pearl, 2013).

a) Design heuristics to identify the possible relations of auxiliar causal PGMs that could be transferred to a target causal PGM.

b) Design a function to estimate the local differences in the probability distributions of target PGM and auxiliar PGM.

c) Design a function that determines if each possible causal relation of auxiliar causal PGM could be transferred to a target PGM. In its design, local differences in probability distribution and the target and auxiliary structures will be considered.

d) Validation of the algorithm. The performance of the model transfer algorithm will be evaluated in its ability for identifying the direction of causal relations of ground truth causal structures. The algorithm will be evaluated following a similar evaluation process to that of stage 2. In this case, for each ground truth causal PGM, a target dataset will be sampled, and auxiliar causal PGMs will be generated. After that, a MEC will be estimated (using the algorithm of stage 2), that then, together with the auxiliar causal PGMs, will be used for finding a target subject-specific causal PGM. Finally, estimated causal PGMs will be compared with the ground truth causal PGMs.

i) Generation of synthetic auxiliar causal PGMs: Adding and deleting edges of the ground truth causal PGMs, the auxiliar causal PGMs will be generated.

  - **ii)** An experiment that contemplates different ground truth causal PGMs and different modification schemes for generating auxiliar causal PGMs will be designed and performed. The normalized structural Hamming distance (NSHD), and orientation precision (Jabbari *et al.*, 2018; Triantafillou & Tsamardinos, 2015), will be used as metrics to evaluate the causal PGMs obtained by the algorithm.

4. **Extension of instance-based transfer algorithm for learning from auxiliary mixed datasets**.

   In this stage, the algorithm developed in stage 2 will be extended for learning the structure of subject-specific causal PGMs, transferring knowledge from auxiliary mixed datasets with observational and experimental samples.

   - **a)** The function designed in step 2 will be extended to consider auxiliary datasets with observational and interventional samples.
   - **b)** The functions proposed in the literature for mixed datasets (Brenner & Sontag, 2013; Cooper & Yoo, 1999; Hauser & Bühlmann, 2012), will be analyzed to select one that could be modified, such as that describe the adjust of candidate causal structure with a combination of target and auxiliary datasets.
   - **c)** Define heuristics to combine the causal PGMs learned independently from each auxiliary dataset. In these heuristics should be considered the relevance of each auxiliary dataset and the differences in the set of intervened variables.
   - **d)** Validation: The performance of the algorithm will be evaluated in its ability for identifying the direction of causal relations of the ground truth causal structures. The algorithm will be evaluated following a similar evaluation process to that of stage 2. In this case, for each ground truth causal PGM, a target dataset will be sampled, and auxiliar causal PGMs will be generated. Next, auxiliar datasets will be sampled after performing manipulation on auxiliar causal PGMs. Target dataset and auxiliary datasets will be used to find a target subject-specific causal PGM. Finally, this estimated causal PGM will be compared with the ground truth causal PGMs.
     - **i)** Generation of synthetic datasets with observational and experimental samples. Considering the theory of perfect interventions proposed by Pearl (2000), the scheme of (Luis *et al.*, 2010) will be extended to generate synthetic datasets with observational and experimental samples for the auxiliary domains. This scheme should contemplate variations in the number of intervened variables.
     - **ii)** An experiment that considers variations on the number of intervened variables in auxiliary domains will be designed and performed to evaluate the algorithm. The same metrics used in stage 3 will be used to evaluate the causal structures obtained by the algorithm. It is planned to compare the obtained performance with that obtained by the algorithm of stage 3.

5. **Design and development of a hybrid knowledge transfer algorithm**.

   In this stage, a hybrid algorithm for learning the structure of subject-specific causal PGMs, combining the transfer of instances of auxiliar datasets and causal relations of

Stage	Task	Specific Objective	RQ
1	Assessment of current solutions for structure learning of causal PGMs	1,2,3	1
2	Design and development of instance-based transfer algorithm for learning Markov equivalence classes	1	1
3	Design and development of model-based transfer learning algorithm	2	2
4	Extension of instance-based transfer algorithm for learning from auxiliary mixed datasets	3	3
5	Design and development of a hybrid knowledge transfer algorithm	4	4
6	Exemplification of the knowledge transfer learning algorithm	5	4

Table 4.1: Relation between stages in the methodology, objectives (Section 1.6) and research questions (Section 1.4).

auxiliary PGMs, will be developed. The algorithms of stages 2, 3, and 4, will be integrated for developing this algorithm.

The performance of the hybrid transfer algorithm will be evaluated in its ability for identifying the direction of causal relations of the ground truth causal structures. An experiment that contemplates the process described in stages 3 and 4, for generating auxiliary datasets and causal PGMs, will be designed and performed. The same metrics used in stage 4 will be used to evaluate the causal structures obtained by the algorithm.

6. **Exemplification of the knowledge transfer learning algorithm**.

   In this stage, the hybrid transfer knowledge algorithm will be applied to a causal analysis problem of the neuroscience domain. An experiment will be designed and performed for collecting the data. In its design, a task that involves a well-known neural circuit will be chosen, so that the causal model obtained by the algorithm can be validated with the causal relations reported in the literature.

Important elements for this methodology are described in figures 4.1 and 4.2. In Figure 4.1, the relation between the main stages of this methodology is described. The general scheme that will be used to validate the algorithms is presented in Figure 4.2.

The relation between stages in the methodology, objectives and research questions is presented in Table 4.1.

4.2 Schedule

In Figure 4.3 is presented the schedule of activities for the realization of this research.

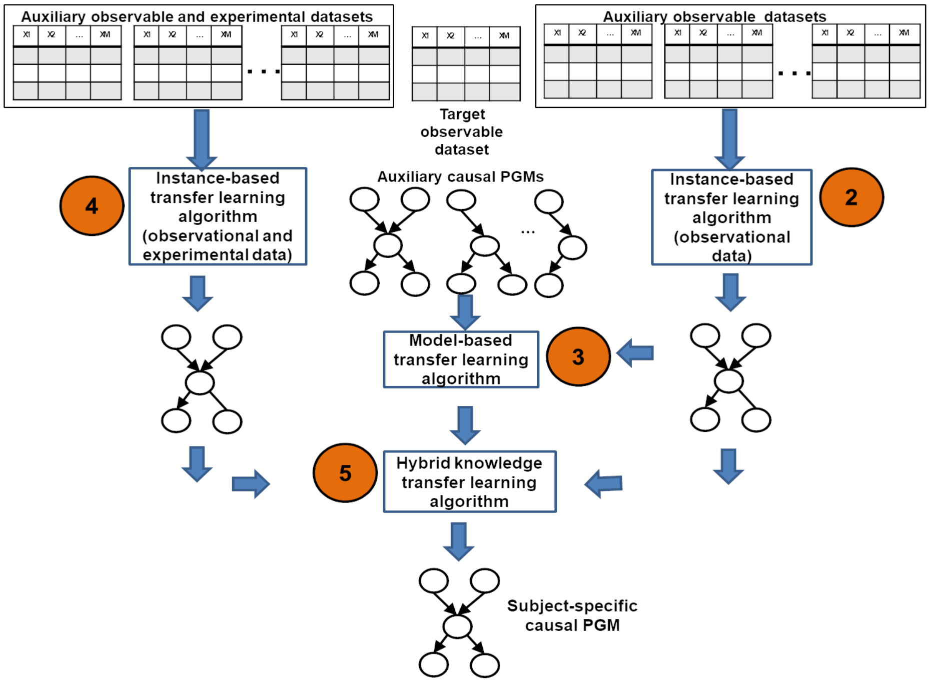

Auxiliary observable and experimental datasets
X1 X2 ... XM
Target observable dataset
Auxiliary observable datasets
4
Instance-based transfer learning algorithm (observational and experimental data)
Auxiliary causal PGMs
...
Instance-based transfer learning algorithm (observational data)
2
Model-based transfer learning algorithm
3
5
Hybrid knowledge transfer learning algorithm
Subject-specific causal PGM

Figure 4.1: Main stages of the methodology: In stage 2, an instance-based transfer algorithm for learning MECs (preliminary subject-specific causal PGMs) will be developed. These preliminary subject-specific causal PGMs will be improved with the model-based transfer algorithm of stage 3. Then, in stage 4, the algorithm designed in stage 2 will be extended for transferring instances from observational and interventional auxiliary datasets. Finally, the algorithms developed in stages 2, 3, and 4, will be integrated into the hybrid knowledge transfer algorithm for learning subject-specific causal PGMs.

4.3 Publications Plan

1. Instance-based transfer for learning Markov equivalence classes. Conference, *The 33rd International Conference of the Florida Artificial Intelligence Research Society*,deadline: November 18, 2019; conference: May 2020.

2. Model-based transfer for learning subject-specific causal PGMs. Conference, *Probabilistic Graphical Models 2020*, deadline: May 2020; conference: September 2020.

3. Instance-based transfer for learning subject-specific causal PGMs. Conference, *Conference on Uncertainty in Artificial Intelligence*, deadline: March 2021; conference: July 2021.

4. Knowledge transfer for learning subject-specific causal PGMs. Journal, *International Journal of Approximate Reasoning*, June 2021.

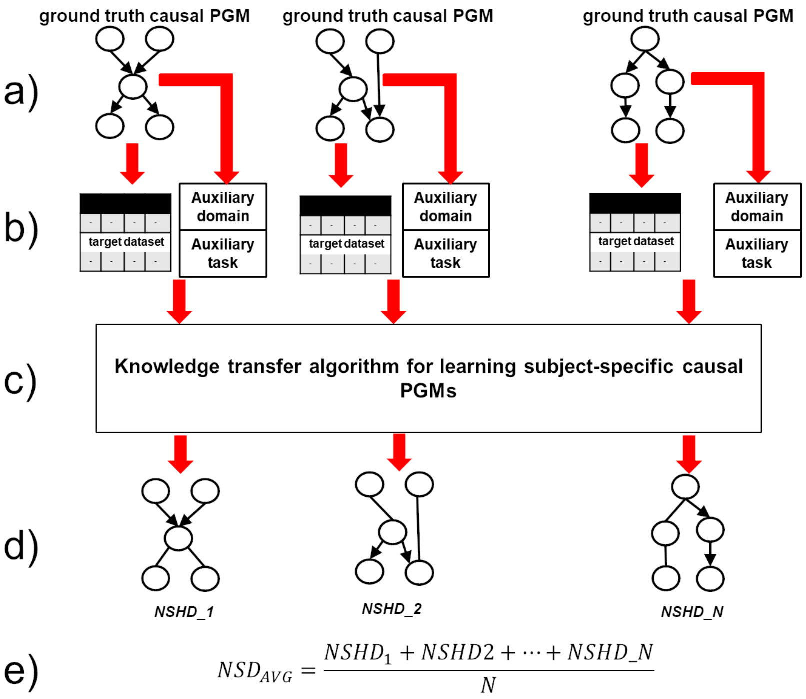
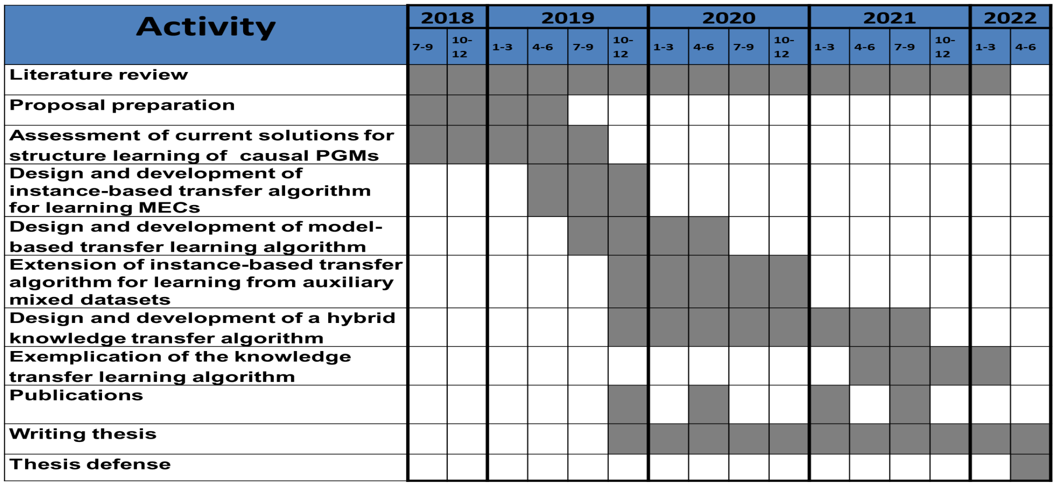


Figure 4.2: Performance evaluation testing for the knowledge transfer learning algorithms of stages 2, 3, 4, and 5. a) Ground truth causal PGMs will be causal PGMs with known structure and parameters. b) From these ground truth causal PGMs will be sampled the target datasets, and the auxiliary domains and tasks will be generated. Each auxiliar task will be composed by a causal PGM, which is a modification of the ground truth causal PGM, and each auxiliary domain, by a dataset sampled from auxiliary tasks. c) The knowledge transfer algorithm for learning subject-specific causal PGMs and d) the subject-specific causal PGMs will be estimated from each target dataset and auxiliary sources. Finally, e) the performance of the algorithm will be obtained from the average of the individual evaluations for each estimated causal structure (exemplified with the normalized structural Hamming distance).



Activity	2018		2019				2020				2021				2022	
	7-9	10-12	1-3	4-6	7-9	10-12	1-3	4-6	7-9	10-12	1-3	4-6	7-9	10-12	1-3	4-6
Literature review	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	
Proposal preparation	■	■	■	■												
Assessment of current solutions for structure learning of causal PGMs	■	■	■	■	■											
Design and development of instance-based transfer algorithm for learning MECs				■	■	■										
Design and development of model-based transfer learning algorithm					■	■	■	■								
Extension of instance-based transfer algorithm for learning from auxiliary mixed datasets						■	■	■	■	■						
Design and development of a hybrid knowledge transfer algorithm						■	■	■	■	■	■	■	■			
Exemplication of the knowledge transfer learning algorithm												■	■	■	■	
Publications						■		■			■		■			
Writing thesis						■	■	■	■	■	■	■	■	■	■	■
Thesis defense																■

Figure 4.3: Schedule of activities for the PhD research.

Chapter 5

Preliminary Results

In this Chapter, preliminary results for this research are presented. These preliminary results are organized in two parts. In the first part, the preliminary results related to the development of an instance-based transfer algorithm for learning Markov equivalence classes are presented. The preliminary results obtained for an experiment of brain functional connectivity analysis are presented in the second part.

These preliminary results are part of the specific objectives first, and five that are related to steps one, two, fourth, and six of the methodology.

5.1 Knowledge transfer-GES

The proposed preliminary algorithm is an extention of the Greedy Equivalence Search (GES) algorithm that from two auxiliary observational datasets, finds the Markov equivalence class (MEC) corresponding to a target subject-specific causal PGM. Under the assumptions of causal sufficiency and faithfulness conditions, the best MEC is found by maximizing a score function that combines the knowledge of target and source domains.

For combining the knowledge of target and source domains, local knowledge transfer of the best source domain is explored. In this local knowledge transfer, the instances of the best source domains are transferred for finding the parents set $\mathbf{Pa}(X_i)$ of each node X_i in a MEC $\mathcal{G}$. The best source domain among the available pool is determined using a fitness function that measures the relatedness of the source domain with the target domain. The proposed local knowledge transfer is applied to each iteration of the GES when a candidate structure is evaluated with the score function, in such form as described in Algorithm 2.

Two new forms of the local knowledge transfer are proposed here: knowledge transfer of the weighted-instances and knowledge transfer of re-sampling instances. These forms of extending the GES algorithm are described in the following sections.

5.1.1 Knowledge transfer of weighted-instances

This extension of GES, denominated as TKL-WeGES, uses the local transfer of weighted instances of auxiliary domains for estimating the score of a candidate MEC in the target domain.

Algorithm 2: SCORETKL algorithm

```
Function scoreTKL()
    Input: $\mathbf{X}$
    $D_T$: The observable datasets for target domain
    $D_{S1}, D_{S2}$: The observable datasets for source domains
    $\mathcal{G}'$: the candidate structure for target
    $\mathcal{S}_1, \mathcal{S}_2$: the structures for source domains
    Output: $s$: the score for $\mathcal{G}'$
    $s \leftarrow 1$
    foreach $X \in \mathbf{V}$ do
        $s \leftarrow s * localScoreTKL(X, \mathbf{Pa}_T(X), \mathbf{Pa}_{S1}(X), \mathbf{Pa}_{S2}(X), D_T, D_S)$
    end
    return $s$
Function localScoreTKL()
    Input: $X$
    $\mathbf{Pa}_T(X)$: parents of $X$ in $\mathcal{G}'$
    $\mathbf{Pa}_{S1}(X)$: parents of $X$ in $\mathcal{G}_{S1}$
    $\mathbf{Pa}_{S2}(X)$: parents of $X$ in $\mathcal{G}_{S2}$
    Output: $s$ score for $\mathcal{G}'$
    $lf_{S1} \leftarrow localFitness(X, D_T, D_{S1}, \mathbf{Pa}_T(X), \mathbf{Pa}_{S1}(X))$
    $lf_{S2} \leftarrow localFitness(X, D_T, D_{S2}, \mathbf{Pa}_T(X), \mathbf{Pa}_{S2}(X))$
    if $(lf_{S2} < lf_{S1})$ then
        $s \leftarrow localScore(X, D_T, D_{S1}, \mathbf{Pa}_T(X), lf_{S1})$
    end
    else
        $s \leftarrow localScore(X, D_T, D_{S2}, \mathbf{Pa}_T(X), lf_{S2})$
    end
    return $s$
```

The local BDeU score defined in the Equation 2.4 is used for evaluating the adjust of the combination of source D_S and target D_T datasets with a local structure composed by $X_i \in (X)$ with their parents $\mathbf{Pa}_T(X_i)$, as follows:

$$localScoreWe(X_i, \mathbf{Pa}_T(X_i), D_T, D_S) = \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + NC_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + NC_{ijk})}{\Gamma(\alpha_{ijk})} \tag{5.1}$$

In this local score, $NC_{ijk} = (N_{ijk})_T + W_i(N_{ijk})_S$ considers the combinations of the target instances with the weighted instances of source domains. $(N_{ijk})_T$ represents the number of cases in D_T in which $X_i = k$ and its parents $\mathbf{pa}_T(X_i) = j$, and $(N_{ijk})_S$, the number of cases in D_S in which $X_i = k$ and its parents $\mathbf{pa}_T(X_i) = j$. Besides W_i, encode the differences in the conditional probability distribution of X_i between the target $P_T(X_i|\mathbf{Pa}_T(Xi))$ and the source domain $P_S(X_i|\mathbf{Pa}_S(Xi))$.

For estimating W_i, the Kullback-Leibler divergence is used as follows:

$$W_i = 2^{-|D_{KLD}(P_T(x_i|\mathbf{pa}_T(X_i)), P_S(x_i|\mathbf{pa}_S(X_i)))|} \tag{5.2}$$

With this function, when the difference between target and auxiliary datasets increases, it is penalized with weights nearly to zero; and it assigns weights nearly to one, to small differences lower to one.

Considering that W_i encodes the difference between the target and the source domains, it is also used to measure their relatedness as follows,

$$localFitness(X_i, D_T, D_S) = W_i \tag{5.3}$$

5.1.2 Knowledge transfer of re-sampling instances

This is an extension of GES that uses the transfer of re-sampling instances of the target domain (denominated as TKL-ReSGES). We assume that new instances for the target domain are generated for estimating the score of each candidate MEC. These new instances for the target domain are sampling from the Dirichlet-Multinomial distribution with *a priori* parameters $(\alpha + (\mathbf{N})_S)$, where $(\mathbf{N})_S$ is the counts of the source domain. The new instances for the target domain have a Dirichlet-Multinomial distribution with parameters $(\alpha + (\mathbf{N})_S + (\mathbf{N})_T)$.

This Dirichlet-Multinomial distribution with parameters $(\alpha + (\mathbf{N})_S + (\mathbf{N})_T)$ is used to estimate the probability that a candidate MEC for a target domain, adjusts with the combination of instances for source and target domains, that is, $P(\mathcal{G}|D_T, D_s) = DIR(\alpha + (\mathbf{N})_S + (\mathbf{N})_T)$.

Considering that, the local score function is defined as follow (Zhou *et al.*, 2016):

$$localScoreRe(X_i, \mathbf{Pa}_T(X_i), D_T, D_S) = \prod_{j=1}^{q_i} \frac{\Gamma((AN_{ij})_S)}{\Gamma((AN_{ij})_S + (N_{ij})_T)} \prod_{k=1}^{r_i} \frac{\Gamma((AN_{ijk})_S + (N_{ijk})_T)}{\Gamma((AN_{ijk})_S)} \tag{5.4}$$

where,

n is the number of nodes in $\mathcal{G}$,

q_i is the number of instantiations of $\mathbf{Pa}(X_i)$,

r_i is the number of values of X_i,

$(AN_{ijk})_S = \alpha_{ijk} + (N_{ijk})_S$ is agregated counts from the prior distribution α and source domains,

$(AN_{ij})_S = \sum_k (AN_{ijk})_S$,

$(N_{ijk})_T$ is the number of cases in D_T in which $X_i = k$ and its parents $\mathbf{pa}_T(X_i) = j$,

$(N_{ij})_T = \sum_k (N_{ijk})_T$,

$\alpha_{ijk} = \frac{\alpha}{r_i q_i}$ is a Dirichlet *prior* parameter.

In this local score, the instances of the best source domain are used. For measuring the relatedness of the source domain with the target domain, is used the local score of equation 2.4 as follows,

$$localFitness(X_i, \mathbf{Pa}_S(Xi), D_T) = localScore(X_i, \mathbf{Pa}(X_i), D_T) \qquad (5.5)$$

5.1.3 Experiments

5.1.3.1 Generation of synthetic datasets

Synthetic datasets are generated from ground truth Bayesian networks which are BN with known structure and parameters. Target datasets and source datasets are generated in the following form. Target dataset is sampled from the ground truth BN, and source datasets, from related BNs. Related BNs are generated from the ground truth BN, adding $pMod$ percent of edges, followed by deleting $pMod$ percent of edges. Next, their parameters are estimated using a dataset sampled from the ground truth BN. Each dataset is sampled from their corresponding BN using forward sampling in which the values of each variable X_i are sampled in ancestral order (parents before their children), in such form that its values x_i are drawn from $P(x_i|\mathbf{pa}(X_i))$. This process was implemented in R using the bnlearn package (Scutari, 2009).

5.1.3.2 Experimental design

In this experiment, we hypothesized that the TKL-WeGES and TKL-ReSGES algorithms would outperform the GES algorithm. The performance of the TKL-WeGES and TKL-ReSGES algorithms was evaluated in their ability for finding skeletons of the ground truth models. In this evaluation, the Coma (Cooper, 1984) and Asia (Lauritzen & Spiegelhalter, 1988) binary BNs with five and eight nodes, respectively, were used as ground truth models. These BNs are shown in Figure 5.1. From each BN, two related BNs, modifying the edges of the original BN in 10% and 40%, were created. Considering extreme cases of relatedness (most and less related) were selected these parameters. Datasets with 1600 and 12800 samples for Coma and Asia were used for estimating their respective parameters. Taking into account that after modifying the ground truth BNs would increase the number of parents for some nodes. The sample size was estimated using $samplesize = 100(2^k)$, considering that a node in a related BN may have at most $k = n - 1$ parents (where n is the number of nodes in the BN). In the experiments, for each source dataset, 1600 samples for Coma and 12800 samples for Asia (using the same formule that for the parameters estimation), were obtained. Ten datasets varying

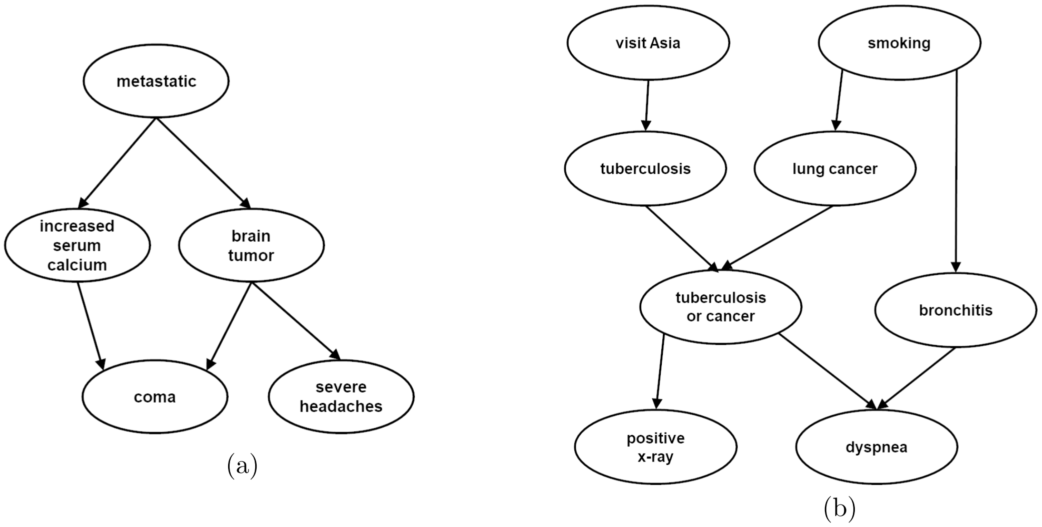


Figure 5.1: Ground truth Bayesian networks for the experiments: a) Coma and b) Asia.

the sample size were obtained for the target domain. For Coma, the set of target datasets includes datasets with size $\{50, 100, 150, 200, 250, 300, 350, 400, 450, 500\}$, and for Asia, with $\{80, 160, 240, 320, 400, 480, 560, 640, 720, 800\}$. Ten runs of this scenary were used to evaluated the algorithms. The models obtained by the algorithms were evaluated using normalized structural Hamming distance (NSHD), adjacency precision (TPR), and adjacency recall (TDR). Normalized structural Hamming distance is the minimum number of edge insertions, deletions, and changes needed to transform a model into another (Montero-Hernandez *et al.*, 2018). Adjacency precision is the ratio $TP/(TP + FP)$, and the ratio $TP/(TP + FN)$ is the adjacency recall. Where TP is the number of adjacencies that are in common in the estimated model and ground truth model without considering the edge orientation; FP is the number of adjacencies that are present in the estimated model but not in the ground truth model; and FN is the number of adjacencies that are present in the ground truth model but not in the estimated model (Jabbari *et al.*, 2018).

5.1.4 Results

The experimental results are summarized in Tables 5.1 and 5.2 for Coma and Asia, respectively. For the TKL-WeGES, the results for transferring instances from the most related domain, and both domains are presented. The results show that TKL-WeGES and TKL-ReSGES improve the models retrieved with respect to GES. In the case of Coma, considering the results for NSHD (the best NSHD is obtained when it is zero), both algorithms seem to decrease the differences between the true and the estimated model. The results for this model also show that, although the performance of the correct edges rate (TPR) seems to decrease, both algorithms are discovering more number of edges, increasing the adjacency discovery of the true model. The results for Asia show an improvement in the TPR and TDR rates. They also

Method	TPR	TDR	NSHD
GES	0.91(0.13)	0.56(0.25)	0.54(0.25)
TKL-WeGES (most related domain)	0.83(0.08)	0.94(0.10)	0.46(0.32)
TKL-WeGES (both domains)	0.83(0.08)	0.94(0.10)	0.54(0.35)
TKL-ReSGES	0.80(0.12)	0.72(0.21)	0.56(0.39)

Table 5.1: Averages of adjacency precision (TPR), adjacency recall (TDR), and normalized structural Hamming distance (NSHD) for Coma. The numbers in parenthesis are standard deviations.

Method	TPR	TDR	NSHD
GES	0.71(0.28)	0.58(0.31)	0.98(0.44)
TKL-WeGES (most related domain)	0.96(0.07)	0.95(0.06)	2.03(0.33)
TKL-WeGES (both domains)	0.89(0.19)	0.90(0.23)	1.89(0.35)
TKL-ReSGES	0.88(0.13)	0.69(0.23)	1.29(0.30)

Table 5.2: Averages of adjacency precision (TPR), adjacency recall (TDR), and normalized structural Hamming distance (NSHD) for Asia. The numbers in parenthesis are standard deviations.

show that the differences between the true and the estimated model increase, which indicate that the estimated model has more edges than the true model (spurious edges).

From the results of both models, it can be observed that both algorithms improve the adjacency discovery rate of the true model (TDR), being better with TKL-WeGES and superior with TKL-ReSGES. Although they also indicate that both algorithms are discovering spurious edges, which seem to increase when the number of nodes increases. It can also be observed that performance for TKL-WeGES is better when the knowledge is transferred from the best domain, than when it is transferred from both domains.

The knowledge transfer of weighted-instances appears to be the most appropriate to be included in a score-based algorithm for discovery MECs. Although, considering the results in NSHD, it is necessary to improve the stage of removing edges of the algorithm.

5.1.5 Conclusions

The advances in the development of a preliminary instance-based transfer algorithm for learning Markov equivalence classes were presented. The advances show the results obtained by analyzing two types of local instances transfer: knowledge transfer of the weighted-instances and knowledge transfer of re-sampling instances. Experimental results indicate that it is feasible to extent a score-based algorithm with the local knowledge transfer of the weighted-instances.

As part of the continuation of this study, we propose to integrate a strategy based on constraints to eliminate the number of incorrect edges. We also consider extending the local

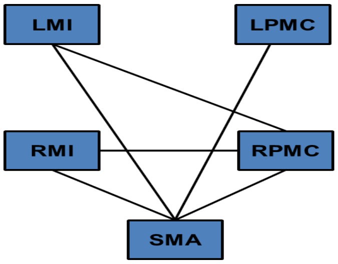


Figure 5.2: Motor execution network: left primary motor area (LM1), right primary motor area (RM1), left pre-motor cortex (LPMC), right pre-motor cortex (RPMC), and supplementary motor area (SMA).

knowledge transfer of the weighted-instances for more than two domains and integrating the Meek orientation rules for deciding some possible orientations of the edges in a MEC (Meek, 2003).

5.2 Functional Connectivity Analysis for Neurorehabilitation Stroke Patients

Functional connectivity refers to the study of the presence of statistical dependencies between specific physiological signals of the brain (Ide *et al.*, 2014). This brain connectivity could be modeled with probabilistic graphical models using data of functional magnetic resonance imaging (fMRI).

Preliminary results for an experiment related to the functional connectivity analysis for patients who followed a Neurorehabilitation therapy are presented. In this experiment, probabilistic graphical models (PGMs) were used for modeling the functional connectivity of stroke patients. In particular, five regions of the motor execution network were considered to the analysis: left primary motor area (LM1), right primary motor area (RM1), left pre-motor cortex (LPMC), right pre-motor cortex (RPMC), and supplementary motor area (SMA). The relation between these regions is described in Figure 5.2 (Bajaj *et al.*, 2015).

5.2.1 Experiment

5.2.1.1 Dataset

A dataset that includes 32 functional magnetic resonance imaging (fMRI) was used in the experiment. This dataset was obtained from eight stroke patients who followed sessions of a virtual reality-based Gesture Therapy. The fMRIs were collected at 4-time points of the therapy: the fMRI scans of the first and the last time point (fMRI-1 and fMRI-4) were taken before to start the therapy and after the end of the planned therapy, respectively. The fMRI for intermediate points (fMRI-2 and fMRI-3) correspond to intermediate therapy sessions (7th and 14th, respectively) (Orihuela-Espina *et al.*, 2013; Orihuela-Espina & Sucar, 2011).

The fMRI images were preprocessed with the Statistical Parametric Mapping (SPM) software (Friston, 2019). All fMRIs were first realigned to correct motion problems. Then, anatomical and functional MRIs were co-registered and spatially normalized to the standard template of the Montreal Neurological Institute (MNI). Finally, they were smoothed using a Gaussian filter of 8mm in width.

After the preprocessing, five regions of the motor execution network were considered to the analysis of brain connectivity: left primary motor area (LM1), right primary motor area (RM1), left pre-motor cortex (LPMC), right pre-motor cortex (RPMC), and supplementary motor area (SMA). The information of each region was obtained using the SPM software, as spheres with 6mm in radius centered at the peak of its corresponding MNI coordinates. The MNI coordinates of each region were defined according to the studies of Bajaj *et al.* (2015); Chen *et al.* (2018). The high pass filtered time-series data of each region were obtained. After that, the data of each region was normalized in the interval [0,1], and discretized in two values. Following this process, the discretized data of all regions for the four fMRIs of each patient were obtained.

5.2.1.2 Experimental design

For this experiment, we hypothesized that there are variations in the probabilistic relations for the motor execution network across the patients and the four sessions of the Neurorehabilitation therapy. The datasets available for seven patients were used in this experiment. Incomplete datasets for the patient six was discarded. GES algorithm was used to find the PGMs encoding the probabilistic relations between the brain regions.

The PGM for each patient and each session was constructed, yielding in a total of 28 models. This scenery was run seven times, leaving out in each run, the datasets for a patient. The most frequent relation for all runs was used to integrated the final PGM.

The difference between models was evaluated using False positive rate (TFR). It is defined as the adjacencies of the one model that not are present in the other model (TF), overall adjacencies of the one model (TP+TF). The averages of the TFR over all comparations between pairs of models, corresponding to the same session, was finally used to evaluate the differences in connectivity between patients. A similar process, using all comparison between the same patient overall sessions, was used to estimate the differences between sessions.

5.2.2 Results

The probabilistic graphical models obtained for six patients are presented in Figures 5.3 and 5.4. These figures show the models obtained by GES for each patient and each session. The PGMs without functional connections for the patient four are not presented.

From these results, we can observe that there are variations in the functional connectivity across patients and sessions, with average (standard deviation) in TFR of 0.24(0.06) and 0.36(0.12) , respectively. The patients with high functional connectivity are those corresponding to patients five and seven, and with low connectivity, patients one, two and six. It can be observed, that the connectivity seems to be more variable across the sessions. For some patients, the connectivity decrease across sessions, and in others the connectivity seems to increase, and decrease in the last session. Moreover, for almost all patients, some relation between regions changes across the sessions.

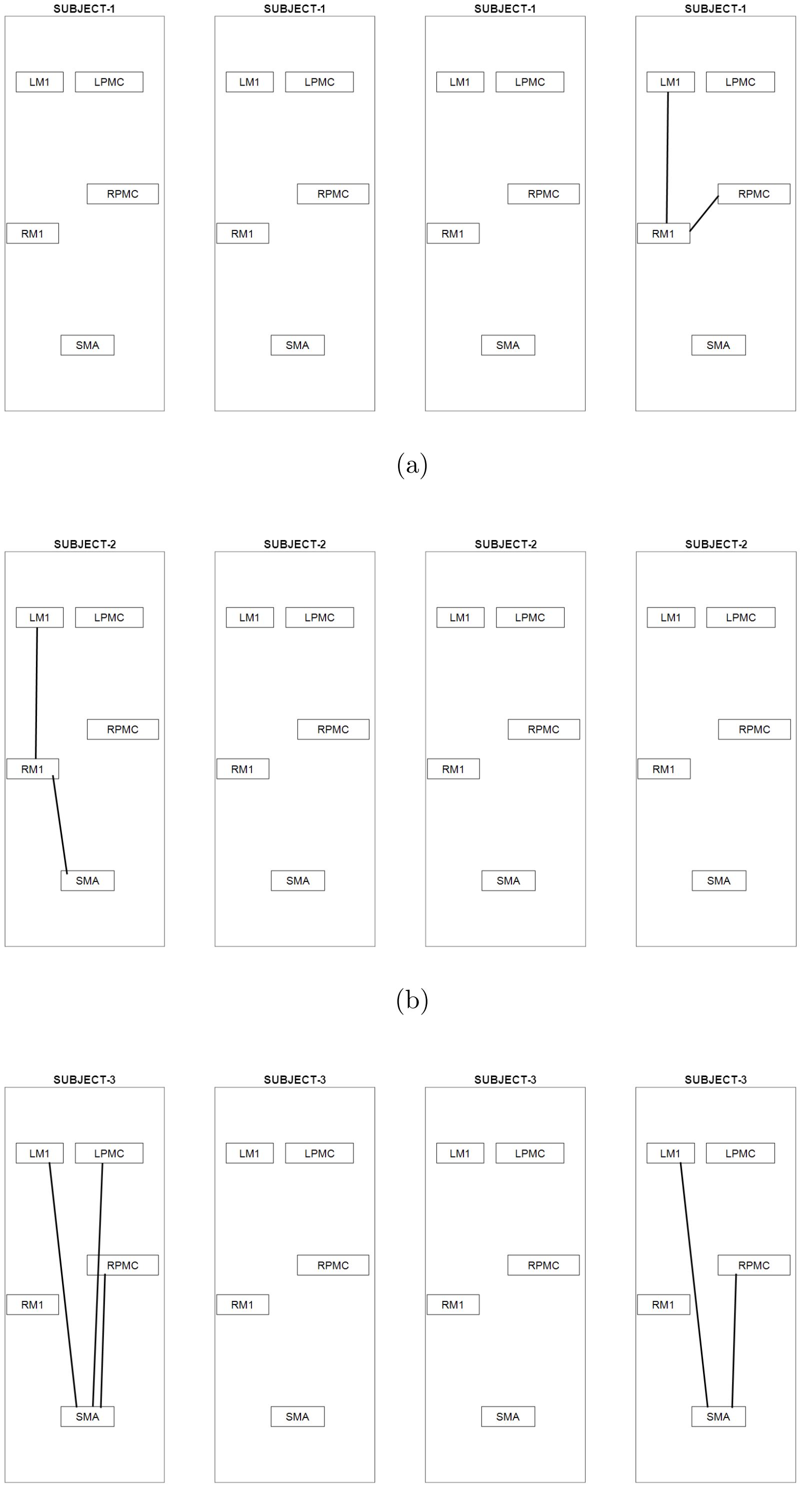
Comparing the models with the motor execution network described in Figure 5.2, the models that seem to be more consistent (with more coincidences) are those corresponding to patients three and seven. It also can be seen that some connections coincide, LMI-SMA, LPMC-SMA, RPMC-SMA, being the most frequents, and RMI-SMA the less frequent.

The results for this experiment show that there are variations between the functional connectivity across the patients and the sessions. It remains to complement the results with other studies that help in their interpretation and validation.

5.2.3 Conclusions

The preliminary results obtained by an experiment of functional connectivity for the motor execution network of patients who followed neurorehabilitation therapy were presented. Preliminary results show that there are variations in the functional connectivity for the motor execution network, across subjects and sessions.

As part of the continuation of this study, we intend to apply the algorithms to be developed in this research, for improving the results, and also for identifying which of the functional connections are part of causal relations.

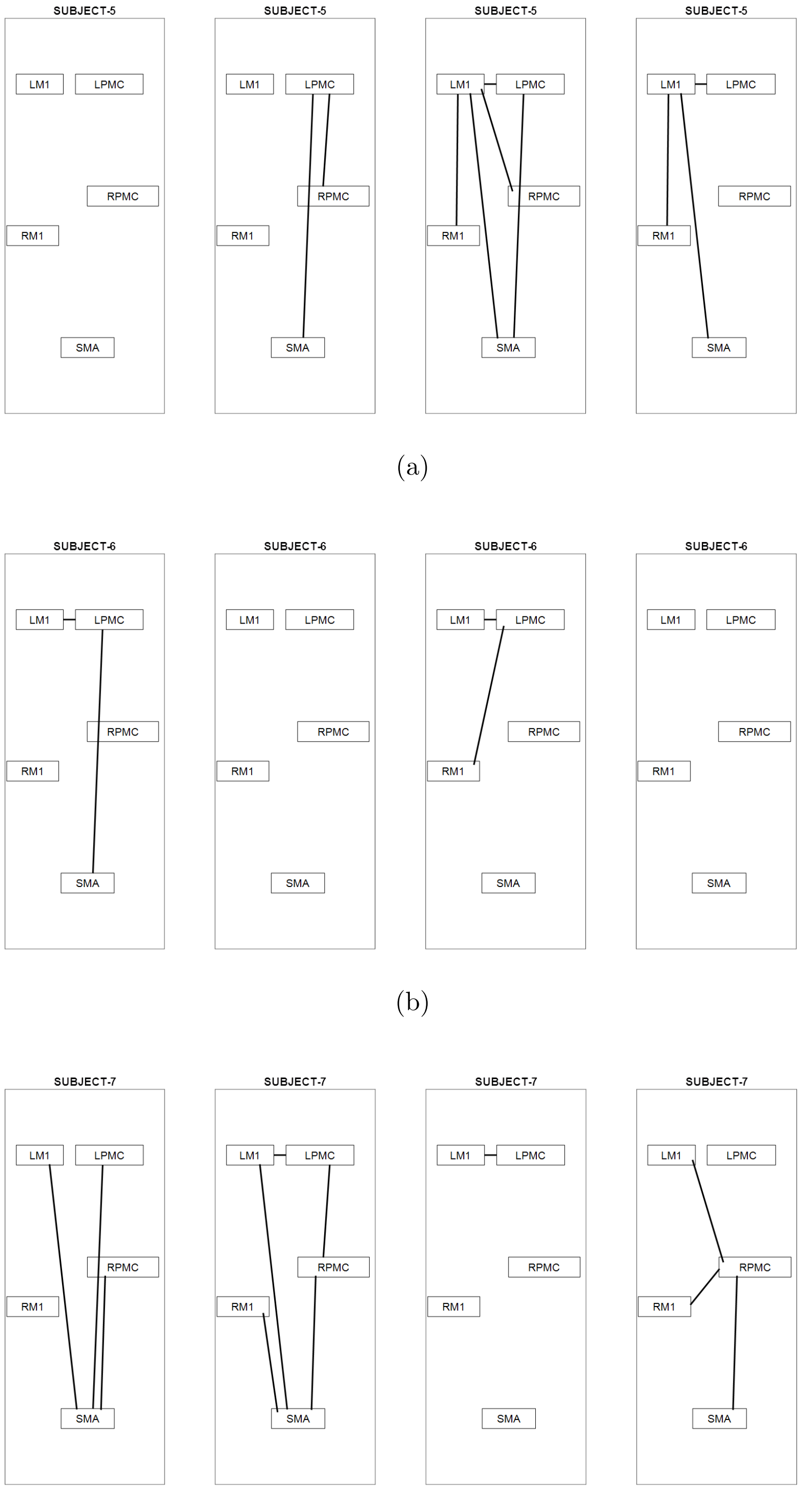


(a)

(b)

(c)

Figure 5.3: Functional conectivity models for patients (a) one, (b) two, and (c) three corresponding to the sessions one to four.



(a)

(b)

(c)

Figure 5.4: Functional conectivity models for patients (a) five, (b) six, and (c) seven, corresponding to the sessions one to four.

Bibliography

Charu C Aggarwal. *Data classification: algorithms and applications*. CRC press, 2014.

Sahil Bajaj, Andrew J Butler, Daniel Drake, and Mukesh Dhamala. Brain effective connectivity during motor-imagery and execution following stroke and rehabilitation. *NeuroImage: Clinical*, 8:572–582, 2015.

Elias Bareinboim and Judea Pearl. Causal transportability with limited experiments. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*, 2013.

Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.

Eliot Brenner and David Sontag. Sparsityboost: A new scoring function for learning Bayesian network structure. *arXiv preprint arXiv:1309.6820*, 2013.

Lindsey Jennifer Fiedler Cameras, Luis Enrique Sucar, and Eduardo F Morales. A transfer learning approach for learning temporal nodes Bayesian networks. In *The Twenty-Sixth International FLAIRS Conference*, 2013.

Luis M de Campos. A scoring function for learning bayesian networks based on mutual information and conditional independence tests. *Journal of Machine Learning Research*, 7(Oct): 2149–2187, 2006.

Jing Chen, Dalong Sun, Yonghui Shi, Wei Jin, Yanbin Wang, Qian Xi, and Chuancheng Ren. Alterations of static functional connectivity and dynamic functional connectivity in motor execution regions after stroke. *Neuroscience Letters*, 686:112–121, 2018.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3(Nov):507–554, 2002.

Tom Claassen and Tom Heskes. Causal discovery in multiple models from different experiments. In *Advances in Neural Information Processing Systems*, pages 415–423, 2010.

Gregory Cooper, Chunhui Cai, and Xinghua Lu. Tumor-specific causal inference (TCI): A Bayesian method for identifying causative genome alterations within individual tumors. *bioRxiv*, page 225631, 2018.

Gregory F Cooper and Changwon Yoo. Causal discovery from a mixture of experimental and observational data. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 116–125. Morgan Kaufmann Publishers Inc., 1999.

Gregory Floyd Cooper. Nestor: A computer-based medical diagnostic aid that integrates causal and probabilistic knowledge. Technical report, Stanford University CA, Dept of Computer Science, 1984.

Karl Friston. Statistical parametric mapping. https://www.fil.ion.ucl.ac.uk/spm/, 2019. Accessed February 2019.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10:1–15, 2019.

Christian Grefkes and Gereon R Fink. Connectivity-based approaches in stroke and recovery of function. *The Lancet Neurology*, 13(2):206–216, 2014.

Ruocheng Guo, Lu Cheng, Jundong Li, P Richard Hahn, and Huan Liu. A survey of learning causality with data: Problems and methods. *arXiv preprint arXiv:1809.09337*, 2018.

Alain Hauser and Peter Bühlmann. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13(Aug):2409–2464, 2012.

Yangbo He, Jinzhu Jia, and Bin Yu. Counting and exploring sizes of Markov equivalence classes of directed acyclic graphs. *The Journal of Machine Learning Research*, 16(1):2589–2609, 2015.

Christina Heinze-Deml, Marloes H Maathuis, and Nicolai Meinshausen. Causal structure learning. *Annual Review of Statistics and its Application*, 5:371–391, 2018.

Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, pages 689–696, 2009.

Jaime S Ide, Sheng Zhang, and R Li Chiang-shan. Bayesian network models in brain functional connectivity analysis. *International Journal of Approximate Reasoning*, 55(1):23–35, 2014.

Fattaneh Jabbari, Shyam Visweswaran, and Gregory F Cooper. Instance-specific Bayesian network structure learning. In *International Conference on Probabilistic Graphical Models*, pages 169–180, 2018.

Haiyang Jia, Zuoxi Wu, Juan Chen, Bingguang Chen, and Sicheng Yao. Causal discovery with Bayesian networks inductive transfer. In *International Conference on Knowledge Science, Engineering and Management*, pages 351–361. Springer, 2018.

Markus Kalisch and Peter Bühlmann. Causal structure learning and inference: a selective review. *Quality Technology & Quantitative Management*, 11(1):3–21, 2014.

Daphne Koller, Nir Friedman, and Francis Bach. *Probabilistic graphical models: principles and techniques.* MIT press, 2009.

Steffen L Lauritzen and David J Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(2):157–194, 1988.

Junning Li, Z Jane Wang, Samantha J Palmer, and Martin J McKeown. Dynamic Bayesian network modeling of fMRI: a comparison of group-analysis methods. *Neuroimage*, 41(2): 398–407, 2008.

Xiang Li, Shanghong Xie, Peter McColgan, Sarah J Tabrizi, Rachael I Scahill, Donglin Zeng, and Yuanjia Wang. Learning subject-specific directed acyclic graphs with mixed effects structural equation models from observational data. *Frontiers in Genetics*, 9, 2018.

Roger Luis, L Enrique Sucar, and Eduardo F Morales. Inductive transfer for learning Bayesian networks. *Machine learning*, 79(1-2):227–255, 2010.

Daniel Malinsky and David Danks. Causal discovery algorithms: A practical guide. *Philosophy Compass*, 13(1):e12470, 2018.

Andrea Mechelli, Will D Penny, Cathy J Price, Darren R Gitelman, and Karl J Friston. Effective connectivity and intersubject variability: using a multisubject network to test differences and commonalities. *Neuroimage*, 17(3):1459–1469, 2002.

Christopher Meek. Complete orientation rules for patterns. Technical report, Carnegie Mellon University, 2003.

Toni Monleón Getino and Jaume Canela i Soler. Causality in medicine and its relationship with the role of statistics. *Biomedical Statistics and Informatics, 2017, vol. 2, num. 2, p. 61-68*, 2017.

Samuel Montero-Hernandez, Felipe Orihuela-Espina, and Luis Enrique Sucar. Intervals of causal effects for learning causal graphical models. In *International Conference on Probabilistic Graphical Models*, pages 296–307, 2018.

Joris M Mooij, Sara Magliacane, and Tom Claassen. Joint causal inference from multiple contexts. *arXiv preprint arXiv:1611.10351*, 2019.

Richard E. Neapolitan. *Learning Bayesian networks*, volume 38. Pearson Prentice Hall Upper Saddle River, NJ, 2004.

Alexandru Niculescu-Mizil and Rich Caruana. Inductive transfer for Bayesian network structure learning. In *Artificial Intelligence and Statistics*, pages 339–346, 2007.

Chris J Oates, Jim Q Smith, Sach Mukherjee, and James Cussens. Exact estimation of multiple directed acyclic graphs. *Statistics and Computing*, 26(4):797–811, 2016.

Juan Miguel Ogarrio, Peter Spirtes, and Joe Ramsey. A hybrid causal search algorithm for latent variable models. In *Conference on Probabilistic Graphical Models*, pages 368–379, 2016.

Felipe Orihuela-Espina, Isabel Fernandez del Castillo, Lorena Palafox, Erick Pasaye, Israel Sánchez-Villavicencio, Ronald Leder, Jorge Hernández Franco, and Luis Enrique Sucar. Neural reorganization accompanying upper limb motor rehabilitation from stroke with virtual reality-based gesture therapy. *Topics in Stroke Rehabilitation*, 20(3):197–209, 2013.

Felipe Orihuela-Espina and Luis Enrique Sucar. Functional reorganization strategies associated to motor rehabilitation gesture therapy. Technical Report 300, Coordinación de Ciencias Computacionales. Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), 2011.

Diane Oyen and Terran Lane. Bayesian discovery of multiple Bayesian networks via transfer learning. In *2013 IEEE 13th International Conference on Data Mining*, pages 577–586. IEEE, 2013.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

Judea Pearl. *Causality: Models, Reasoning, and Inference.* Cambridge University Press, New York, NY, USA, 2000.

Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect.* Basic Books, 2018.

Joseph Ramsey, Madelyn Glymour, Ruben Sanchez-Romero, and Clark Glymour. A million variables and more: the Fast Greedy Equivalence Search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *International Journal of Data Science and Analytics*, 3(2):121–129, 2017.

Joseph D Ramsey, Stephen José Hanson, Catherine Hanson, Yaroslav O Halchenko, Russell A Poldrack, and Clark Glymour. Six problems for causal inference from fMRI. *Neuroimage*, 49(2):1545–1558, 2010.

Marco Scutari. Learning Bayesian networks with the bnlearn R package. *arXiv preprint arXiv:0908.3817*, 2009.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(Oct):2003–2030, 2006.

Peter Spirtes and Clark Glymour. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9(1):62–72, 1991.

Peter Spirtes, Clark N Glymour, Richard Scheines, David Heckerman, Christopher Meek, Gregory Cooper, and Thomas Richardson. *Causation, prediction, and search.* MIT press, 2000.

Peter Spirtes, Christopher Meek, and Thomas Richardson. Causal inference in the presence of latent variables and selection bias. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 499–506. Morgan Kaufmann Publishers Inc., 1995.

Peter Spirtes and Kun Zhang. Causal discovery and inference: Concepts and recent methodological advances. In *Applied informatics*, volume 3, pages 1–28. SpringerOpen, 2016.

Luis Enrique Sucar. *Probabilistic Graphical Models.* Advances in Computer Vision and Pattern Recognition; Springer: London, UK, 2015.

Robert Tillman and Peter Spirtes. Learning equivalence classes of acyclic models with latent and selection variables from multiple datasets with overlapping variables. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 3–15, 2011.

Robert E Tillman and Frederick Eberhardt. Learning causal structure from multiple datasets with similar variable sets. *Behaviormetrika*, 41(1):41–64, 2014.

Sofia Triantafillou and Ioannis Tsamardinos. Constraint-based causal discovery from multiple interventions over overlapping variable sets. *Journal of Machine Learning Research*, 16: 2147–2205, 2015.

Shyam Visweswaran and Gregory F Cooper. Learning instance-specific predictive models. *Journal of Machine Learning Research*, 11(Dec):3333–3369, 2010.

Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big Data*, 3(1):9, 2016.

Tao Wu, Liang Wang, Mark Hallett, Yi Chen, Kuncheng Li, and Piu Chan. Effective connectivity of brain networks during self-initiated movement in parkinson's disease. *Neuroimage*, 55(1):204–215, 2011.

Jiji Zhang and Peter Spirtes. Detection of unfaithfulness and robust causal inference. *Minds and Machines*, 18(2):239–271, 2008.

Jiji Zhang and Peter Spirtes. The three faces of faithfulness. *Synthese*, 193(4):1011–1027, 2016.

Kun Zhang and Aapo Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 647–655. AUAI Press, 2009.

Kun Zhang, Bernhard Schölkopf, Peter Spirtes, and Clark Glymour. Learning causality and causality-related learning: some recent progress. *National Science Review*, 5(1):26–29, 2018.

Yun Zhou, Timothy M Hospedales, and Norman Fenton. When and where to transfer for Bayesian network parameter learning. *Expert systems with applications*, 55:361–373, 2016.

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