Tutorial on Causal Reasoning and Discovery

Synopsis: This tutorial describes the elementary structures and operations of graphical causal models, including causal reasoning and causal discovery. Additionally, it demonstrates the software tools for causal discovery at an introductory level.

Duration: 8 hours

Capacity: 50

Level: Introductory

Instructors:
Adnan Darwiche, Computer Science Dept., University of California Los Angeles (UCLA)
Luis Enrique Sucar, Computer Science Dept., Instituto Nacional de Astrofisica, Optica y Electronica (INAOE)
Samuel Montero-Hernandez, Bio Optical & Acoustic Spectroscopy Lab (BOAS Lab). Boston University.
Julio César Muñoz-Benitez, Computer Science Dept., Instituto Nacional de Astrofisica, Optica y Electronica (INAOE).

Tutorial structure.
The tutorial is divided into four parts:

- Part I. Introduction to Causal Reasoning – A. Darwiche
- Part II. Introduction to Causal Discovery – E. Sucar
- Part III. Tools for Causal Discovery – S. Montero
- Part IV. Tools for Causal Discovery in Time Series – J. Muñoz

Parts I and II will present the fundamentals of causal reasoning and causal discovery

Parts III and IV will demonstrate the application of those concepts through the use of software tools for causal discovery.

Part I. Introduction to Causal Reasoning

I will discuss in this tutorial some fundamentals of the causal hierarchy which embodies a reasoning hierarchy and an information hierarchy. The reasoning hierarchy contains three classes of queries: associational, interventional and counterfactual. The information hierarchy contains models needed for answering these queries: Bayesian networks, causal Bayesian networks, and functional Bayesian networks (also known as structural causal models, SCMs). One focus of the tutorial will be on answering casual queries based on causal graphs and data which are usually not enough to learn a causal model but may be sufficient to learn (i.e., identify) a particular query.

Part II. Introduction to Causal Discovery

Learning causal models from observational data is challenging, as in general we can not recover a unique causal model, but a set of statistically equivalent models that are called a Markov equivalence class. In this tutorial we will cover the fundamentals of causal discovery, including the graph representation of equivalence classes. We will introduce some of the algorithms that can be applied to learning a causal model, and a technique that can try to obtain the most probable model within a Markov equivalence class. Finally we will present some applications, including discovering the effective connectivity in the brain, and learning causal models from COVID data.

Part III. Tools for causal discovery.

The main focus in many scientific fields is understanding the cause-and-effect relationships between variables, often accomplished through experiments. However, conducting experiments may not always be practical due to factors such as time, cost, or ethical concerns. In these cases, the challenge is to determine causal information from observational data. Algorithms such as PC, FCI, and RFCI can infer the causal structure from observational data under certain assumptions, but they do not provide information on the magnitude of causal effects. In this case and under certain assumptions, the IDA method can infer bounds on causal effects using observational data.

To make these methods more accessible, the R package pcalg was created, which includes implementations of the PC, FCI, RFCI algorithms, and the IDA method. The purpose of this tutorial is to introduce the pcalg package, its range of functions, and demonstrate its use with simulated and real data sets.

Learning objectives.
It is essential to employ well-documented and user-friendly software for the broad application of these methods. We will explore the R package pcalg, which includes the implementations of the PC, FCI, and RFCI algorithms, and the IDA method. This tutorial aims to provide a hands-on introduction to the pcalg package and its functionalities using simulated and real datasets.

1. First steps in R
2. Graphical models representation in R
   a. dags
   b. Conditional independence tests
   c. pcalg package
3. Estimating causal structures with graphical models
   a. skeleton
   b. pc
   c. fci
   d. rfci
4. Estimating bounds on causal effects
   a. ida
   b. idaFast

Requirements.
Attendants may optionally bring a laptop with the following tools:

1. R (≥ 3.5.0): http://lib.stat.cmu.edu/R/CRAN/
2. RStudio (Version: 2023.03.1+446), https://posit.co/download/rstudio-desktop/
3. pcalg R package https://cran.r-project.org/web/packages/pcalg/index.html

Tutorial materials.
The materials to be used during the tutorial (slides, tools, data sets, and scripts) can be found at CaDis2023_materials.

Part IV. Tools for Causal Discovery in Time Series

Time series data have served as the basis for causal discovery in various fields of science. In this sense, data collected can provide very precise measurements at regular points of time. One of the main advantages of using time series is that the temporal order of the information can simplify the causal analysis; the order can help to identify the cause and effect relations. That is, the causal driver can be identified as the variable that occurred first. This contains valuable information for determining the causal relationships of the system being analyzed.

In order to model dynamical systems one may use graphical models such as Directed Acyclic Graphs (DAGs), which consist of a series of nodes connected through edges or links directed from parent nodes to child nodes. The nodes in the DAG represent the variables and the links indicate the causal relationships between these variables. In this way, the use of graphical models may serve as a way to understand dynamic events and analyze the causal relations among data.

Learning objectives:

This tutorial is focused on presenting one of the causal analysis tools in time series. *Tigramite* is a causal inference for time series python package. It allows to estimate causal graphs from time series datasets and to use graphs for robust forecasting, estimation and prediction of direct, total, and mediated effects. Causal discovery is based on linear as well as non-parametric conditional independence tests applicable to discrete or continuous time series. In a basic way, we will introduce the use of *Tigramite* for causal discovery in time series.

1. Overview and Introduction to Tigramite
2. Basic usage
    a. Time series model
    b. Plotting data
    c. PCMCI for causal discovery
3. Causal structure in time series
    a. Plotting a causal structure
4. Integrating (expert) assumptions about links
5. Causal effect estimation

Requirements.
Attendants may optionally bring a laptop with the following tools:

1. Google Collab Account
2. Optionally python=3.7/3.8/3.9/3.10
3. We suggest to create a virtual environment and install the requirements
    a. Requirements are listed in requirements.txt at this link

Tutorial materials.
The materials to be used during the tutorial (slides, tools, data sets, and scripts) can be found at Tools for Causal Discovery in Time Series.