DATA CHALLENGE 2: **CAUSALITY**
Deadline: 26th April

A. LEARNING CAUSE-EFFECT IN TIME SERIES DATA

A time series is a collection of values of a given observable at different time points $\{x(t)\}_{t\in T}$. If the underlying process is random, as in A1, it can be more generally viewed as a realization of the corresponding stochastic process $\{X(t)\}_{t\in T}$.

A1 - MODELING THE FINANCIAL MARKET. As an example, consider the following equation

$$X_{t+dt} = X_t \left(1 + rdt + \sigma\mathcal{N}(0,1)\sqrt{dt}\right), \tag{1}$$

known as the *Black-Scholes equation* and widely used in Finance to model the random evolution in time of the price of a stock with risk-free interest rate r and volatility σ.

- Use the Black-Scholes equation to simulate different trajectories for the price of a stock with the same interest rate and volatility of the S&P500 stock in a given time window. How do the trajectories compare with the real behavior of the S&P500?
- Define a stochastic process $\{Y_t\}_{t\in T}$ producing time series that are correlated with the ones from $\{X_t\}_{t\in T}$. Plot two realizations, one from $\{Y_t\}_{t\in T}$ and one from $\{X_t\}_{t\in T}$, on the same plot. Do they seem correlated?
- Compute the lagged cross-correlation of the two time-series in both directions. By looking at the results obtained, can you conclude the existence of a causal link in one of the two direction? Is your conclusion consistent with what you have expected?

LAGGED CROSS-CORRELATION TEST

1. Make sure both time series are stationary (A common test for stationarity is the Augmented Dickey-Fuller Test). If the time series are not stationary, make them stationary.
2. Compute the lagged cross-correlation $\rho(\tau) \equiv \langle\langle X(t-\tau)Y(t)\rangle\rangle$ for both positive (X preceeds Y) and negative (Y preceeds X) values of the lag τ in a range $[-\tau_{max}, +\tau_{max}]$.
3. Visualize the scatter plot $\rho(\tau)$ VS τ. Infer the direction of the causal link from the sign of the lag corresponding to the maximum value of the correlation. In other words, infer X causes Y if the maximum is observed for positive lags (when X preceeds Y) and vice-versa. Take the value of the maximum as an indication of the strenght of the causal link.

A2 - CHAOTIC SYSTEMS. Consider now the following system of equations

$$\begin{cases} X(t+1) = X(t)\left[r_X - r_X X(t) - \beta_{XY} Y(t)\right] \\ Y(t+1) = Y(t)\left[r_Y - r_Y Y(t) - \beta_{YX} X(t)\right] \end{cases} \tag{2}$$

where $r_X = 3.8, r_Y = 3.5, \beta_{YX} = 0.1, \beta_{XY} = 0.02, X_0 = 0.4, Y_0 = 0.2$.

- Simulate the system for $n = 1000$ timepoints and visualize the trajectories in evenly spaced time intervals (e.g. $[0, 100], [100, 200], \dots$). By just looking at the trajectories in different time windows, what can be said about the nature of the system? What would you expect a cross-correlation analysis will result in?

- Perform a cross-correlation analysis and compare the results with your expectations. Would you trust a cross-correlation analysis in this case? In other words, based on the nature of the system, do you think a cross-correlation analysis is telling us something about causality in this system? Why? **If not** Find (or develop by yourself) better causality inference methods and apply them to this case. Comment the results.
- Apply these new causal inference methods to the time series built in A1. Are the results consistent with what you have concluded before?

A3 - CHALLENGE ON REAL DATA (*Who causes who?*). Your task is to apply causal inference methods (which you have either found on the interned or developed by yourself) to infer who causes who in at least **two** pairs of real time series of your choice (which you think might be causally-linked). Describe the methods chosen/developed and discuss their validity (e.g. why do you think the methods are suitable to infer causality in the data you have chosen). Compare your methods with a cross-correlation analysis. Present and discuss your results. For example, are your findings consistent with your expectations? Provide a possible rationale/mechanism/explanation for the inferred causality.

Possible time series data (list not exhaustive - be creative!):

- *Finance.* Time series data in the financial world are, for example, the historical records of a stock's price. One can get this kind of data from yahoo-finance in a .csv format.
- *Google Trends.* Google trends gives you access to what people are/have been searching for in a given time window. For example, during the Covid Lockdown, were queries about home-made pizza and covid19 news correlated?
- *Climate.* An example is the JenaClimate-2009-2016 Dataset, consisting of timeseries recorded at the Weather Station at the Max Planck Institute for Biogeochemistry in Jena, Germany from 2009 to 2016. It contains 14 different quantities (such as air temperature, atmospheric pressure, humidity, wind direction, and so on) recorded every 10 minutes.
- *Air quality.* The Air quality dataset contains hourly averaged measurements (such as temperature, humidity levels, CO concentration, ...) obtained from an Air Quality Chemical Multisensor Device located on the field in a significantly polluted area, at road level,within an Italian city.
- *Climate change.* https://climate.nasa.gov allows you to download timeseries of global CO2, temperature change, and Arctic sea ice area data from NASA.
- Others ...

(BONUS) CAUSAL RESOLUTION OF PARADOXICAL DATA.

THE MONTY HALL GAME. Consider the following well known problem:

"Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, ***who knows what's behind the doors****, opens another door, say No. 3, which has a goat. He then says to you: "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?"*

- By Simulating a large number of Monty Hall games, compare the performance of the following two players: (1) Player 1 never changes the door; (2) Player 2 always changes the door. Plot the performance of each player (evaluated as the fraction of won games) as a function of the total number of games played. What is the most efficient strategy?
- Do the same analysis of the previous point on a slightly modified version of the game in which the host instead **opens a door at random**. What is the most efficient strategy?
- Explain the differences observed in the two version of the games, as well as why these results seem paradoxical, taking advantage of causal models. *[Hint: keep it simple when making causal models. Consider just the three variables O1: door chosen, O2: door opened, O3: door with car. How are these variables causally linked in the original version of the game? How are they causally linked in the modified version?]*