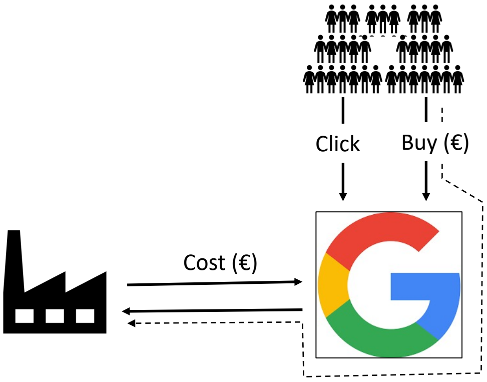
Homework #3, AA2024-2025: Causal impact on e-marketing data (Google) and non-stationary processes

In e-marketing the evaluation of the benefit from digital advertising is evaluated in term of cost and benefit tradeoff. This means that, as a result of some actions, one expects some benefits in terms of a set of clicks and/or selling. These are referred as Key Performance Indicators (KPI), that are numerical indexes to quantify, in this case, the beenfits of a marketing campaign.



The figure illustrates the set of actions in e-marketing. One specific company requires an advertising campain to Google paying a fee (**cost**), and Google sets an online advertising campaing. Google is enabled to monitor anonymously the set of actions by the customers: the interest to a certai advertising (**click**), up to a **buy** of the product. The three set of variables **cost**, **click** and **buy** are typically called Key Performance Indicators (KPIs). The monitoring of the 3 KPIs before or after the beginning of the advertising should have a change of the stochastic properties. To strenghten the quantitative evaluation of the benefit of the advertising, we will extent the usage of the KPIs from other affine products referred as **control** KPIs.

- o -

Let the individual evolution of one KPI (**cost** or **click** or **buy**) vs time for the kth company be $x_k(t)$, and the number of company tracks be $k = 1, 2, ..., K = 20$, where t is a discrete variable that denotes the time samples in days (day 1, day 2, ...). Each of the KPI track is affected by seasonal activity, e.g., summer or winter is more or less active. Furthermore, each trace can change its behavior compared to others as consequence of marketing instances (e.g., product or brand promotion, etc) occurring at the time T_k, the behavour of $x_k(t)$ is a combination of these instances:

$$x_k(t) = \bar{x}_k(t) + g_k(t - T_k)$$

where $\bar{x}_k(t)$ represents the KPI without any marketing istances (i.e., no advertising, no Google adv, etc..), and the behavior $g_k(t)$ is a causal signature, dependent on the specific instance. Signatures are typically fluctuating with a linear growing such as $g_k(t) \simeq \alpha_k t$ for a certain time interval after the instance in T_k. The marketing instances make each signature $x_k(t)$ be non stationary random sequence, even if $\bar{x}_k(t)$ can be considered as stationary.

The tracks $x_0(t), x_1(t), ..., x_K(t)$ are somewhat mutually correlated as likely selected to belong to same retail sectors. This means that, on the average, one can state that $E[x_k(t)x_\ell(\tau)] \neq 0$, and thus the values of each tracks are predictable from the other up to a certain degree of confidence.

Data can be ordered into $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_K]$ of dimension $T \times K$ (here $K = 20$), where each KPI time series are columnwise ordered $T \times 1$:

$$\mathbf{x}_k = \begin{bmatrix} x_k(t=1) \\ x_k(t=2) \\ \vdots \\ x_k(t=T) \end{bmatrix}.$$

Different KPI are active, such as $\mathbf{X}$ (cost for advertising), $\mathbf{Y}$ (clicks), and $\mathbf{Z}$ (conversion: how many customer actions). Notice that numerical values of costs and conversions are are not exactly in € but their values are properly scaled by unknown values. Hw is divided into two parts, Part 1 is on Google dataset, and Part 2 is on detection of discontinuity to detect the excitation time T_k.

Part 1 (Google dataset):

Goal is to consider $x_K(t)$ as track of interest (test time series), and all the other tracks $(k < K)$ are called control time series. The control time series (or simply the control). Notice that in questions 1,2, the KPI $x_k(t)$ denotes cost, or clicks, or conversions (=all together).

1) Derive an estimator of every value $\hat{x}_k(t)$ using linear prediction from $x_k(t)$, by increasing the prediction length and the prediction step. Evaluate quantitatively their properties recalling that for the track of interest $x_K(t)$ for $K = 20$ one should expect a strong variation for $t > T_K = T_{20}$. Propose a metric for performance illustration (what and how to plot, what values, etc..) that details the accuracy of the estimators, and comment the results.

2) Consider the problem of making a set of instantaneous linear estimators $f[.]$ to estimate (*recall that instantaneous means that the estimated valued depends on all other tracks at the same time*)

$$\hat{x}_i(t) = f_i[x_1(t), ...x_k(t), ..., x_K(t)|k \neq i]$$

that is applied to all Ktraces to get $\hat{x}_1(t), \hat{x}_2(t), ..., \hat{x}_K(t)$

2a) define the estimators $f_i[.]$,for any i, and illustrate by a proper metrics to evaluate their accuracy, illustrate and comment the results;

2b) are all control traces useful to mame an accurate estimation? Spot if any control trace can be removed gaining an improvement of the estimated KPI (use each term in the estimator: cost or clicks or conversions);

2c) evaluate if one can estimate the value of advertisinng T_K (assuming not known) from the analysis of the prediction error, illustrate how and with what accuracy.

3) repeat the exercises from 2a,2b,2c) (instantaneous linear estimator) when using all KPI at the same time (cost, and clicks, and conversions).

4) repeat the exercisess from 2a, 2b, 2c) using memory estimators (could be causals, anticausals, or both) and all KPI at the same time (cost, and clicks, and conversions) with unknown length to be defined, and evaluate the configuration that minimize the MSE till the time$T_K = T_{20}$, or using data $t < T_K = T_{20}$

Part 2 (detection of non-stationary processes):

Goal is to design a method to detect when there is a discontinuity as for before and after T_k,so let us consider the following problem: $x[n]$is characterized by a discontinuity in the statistical properties of the a random process to be non-stationary such that:

$$x[n] = \begin{cases} \alpha[n] & n \leq T_o \\ \beta[n] & n > T_o \end{cases}$$

where $\alpha[n]$ is an AR(1) with a pole in $(\rho, 0)$,and $\beta[n]$ is AR(2) with poles in $\rho/2 \cdot \exp(\pm j\phi)$, both AR random processes have power σ_α^2 and σ_β^2, respectively. Goal is to make a classifier of the discontinuity such as before and after the discontinuity the two distributions are different, and thus the corresponding pdf. For the following cases, goal is to evaluate numerically the MSE of the discontinuity position (say $\hat{T}_o$for the following two cases:

1) assume that discontinuity classifier knows $\rho = .8, \phi = \pi/4$, evaluate the a-posteriori probability of the classification for a test-delay $\hat{T}$ sliding over the data such as before $(n < \hat{T})$ the process is AR(1) $\alpha[n]$ and after $(n > \hat{T})$ is AR(2) $\beta[n]$, plot the a-posteriori vs $\hat{T}$ and select the largest for MAP, repeat multiple times as for Montecarlo and evaluate the accuracy of the classifier (e.g, the MSE) that detect the discontinuity time T_o;

2) assume that discontinuity classifier does not know the two AR processes, repeat point 1);

3) use a Neural Network with one hidden layer trained with a set of learning stages (to be defined) to detect the discontinuity T_o as for 1);

4) assume that a classifier knows parameters as in 1) but one want to analyze different configurations vs `rho=linspace(.1,.99,20)`, after following the procedure as in 1) plot now MSE vs ρ.

> *Recall that, as for any Hw, no built-in code can be used and no Matlab packages, otherwise Hw is considered null.*

Matlab GoogleDatset is composed by 3 table, data matrix `Cost`, `Click`, and `Conversion` is ordered in progressive time, and the **control KPI** traces are 1:19 columns (`Cost(:,1:19)`, `Click(:,1:19)`, `Conversion(:,1:19)`), and test of interest is the column 20 (more specifically `Cost(:,20)`, `Click(:,20)`, `Conversion(:,20)`). The instance time from data is $T_{20} = 397$ sample.

-