Mathematics Of Data Curation For Financial Applications

How To Formalize Interactive Bias Discovery And Merger Of Datasets

Charles-Albert Lehalle

Visiting Professor, Imperial College London, UK
Fellow of Institut Louis Bachelier, Paris, France
Global Head of Quantitative R&D Abu Dhabi Investment Authority (ADIA), UAE

Disclaimer

> In these slides, the author expresses his sole opinion and not the one of any of these institutions.

Bachelier Finance Society One World Seminar, May 26, 2022

Outline

Positioning: Datasets? Specific Difficulties In Finance?

From Financial Data To Alternative Data

Financial data are not stationary simply because the economic world is not: international relations, business models, habits, regulations, etc. change. The financial system itself has its own habits and regulations, and the evolution of financial instruments (from linear to always more sophisticated in different directions) produces **a changing environment, shrouded in a mean field of liquidity**.

Market participants try to identify local areas of stationarity via changes of variables (returns are "*more stationary*" than prices; excess returns and cross-sections over well-chosen pools of instruments are more stationary than raw returns); nevertheless **economic objects exhibit memory** (simply because a physical inertia of economic entities), **as well as objects form the financial system**.

But some new datasets (known as **Alternative Data**) exhibit far more stationarity:

- Using satellite images to estimate the biomass under each pixel of daily images of the globe exploit the way **quality of corn is caused by the way its mass grow**.
- Using the supply-chain between factories and car manufacturers exploit how **the capability of building, end hence selling, more cars is caused by the availability of its parts**.

ALTERNATIVE DATA STACK

alternativedata.org

INSTITUTIONAL INVESTORS

ANALYSIS

CLEANING

COLLECTION

yipitDATA

TH DATA CAPITAL

Thinknum

SELL SIDE

Morgan Stanley AlphaWise, UBS Evidence Lab, Data Works, COWEN Data Science Initiative, Bank of America Merrill Lynch, Jefferies

DATA BROKERS

1010data, 7PARK, Pacific Epoch, earnest, Second Measure, alphaMASS insights, Quandl, M | science, SANDALWOOD, TXN, Quantcube

CONSULTANTS

Eagle Alpha, neudata, AVISO, SYSTEM2, ALTERNATIVE DATA GROUP, TOMAHAWK

APP USAGE

App Annie, TUTELA, TalkingData, Mobile Data Value, Mobile

WEATHER

UNDERSTORY, wx, aclima, THE CLIMATE CORPORATION

INFRASTRUCTURE / INTERFACE

Sentieo, CRUX, DOMINO DATA LAB, Adaptive management

WEB DATA

sawr, VERTICAL KNOWLEDGE, Quad analytix, PROMETEIA

CREDIT / DEBIT CARD

ENVESTNET YODLEE, arm INSIGHT, affinity solutions, First Data., MasterCard Advisors, UnionPay 银联, cardlytics

EMAIL RECEIPT

slice intelligence, Superfly insights, Return Path

WEB TRAFFIC

jumpshot, comScore, SimilarWeb, Alexa, AddThis

GEO-LOCATION

Placed, Sense 360, PlaceIQ, THASOS, foursquare, GroundTruth, ADVAN, CUEBIQ, factual.

SATELLITE

Orbital Insight, spire, DigitalGlobe, Descartes Labs, RS Metrics, Ursa, Rezatec, BLACKSKY, tellusLabs, SPACEKNOW, SKYWATCH, SATELLITE IMAGING CORPORATION

SENTIMENT / PUBLIC DATA

Dataminr, visible alpha, RavenPack, estimize, SELERITY, PLURIBUS LABS, accern, DATASIFT, prattle, enigma, hedgyl, iSENTIUM, DISCERN, Sentifi, They Say, MarketPsych, TRITON, Alexandria TECHNOLOGY, socialalpha, LEXALYTICS, Alphamatician

OTHER

comlinkdata, placemeter, PREMISE, GENSCAPE, RIGUP, MarineTraffic, Black Box, str, Receipt Pal, RECEIPT HOG

DATA OWNERS

Alternative Datasets Have Mixed And Unknown Interventions i

Alternative datasets are rarely built for a usage by market participants; they stem from improvements of logistics, marketing or business intelligence purposes. Typically

- Airlines or hotel booking helps travel agents to operate,
- Credit Card datasets mainly helps banks and budget Apps to target the best clients,
- Job-posting databases help employees to chose their next job and companies to adjust their offers.

They can hence often been considered as i.i.d. samples of the correct distribution from the viewpoint of their primary users, **but not for a usage on financial markets**:

- Coming from budget Apps, Credit Card datasets have younger and weathlier people than the clients of retail facing companies,
- Larger companies have more need to advertise their job offers than smaller ones.

Alternative Datasets Have Mixed And Unknown Interventions ii

All these datasets are **partial and come from different sources having different collection process and conditions of collections, that are all partially unknown**.

Seen from a standard statistical viewpoint: **these datasets have biases**, and need to be post-stratified (in the language of *Survey Theory*). The correct variable to be used for this post-stratification have to be discovered in a trial-and-error way, that can be defined as **active post-stratification**.

It raises the question of merging them. You observe $\mathbb{P}(X, Y)$, with no a priori on what Y and X are, and you try to model it, then:

1. **Semi-Supervised Learning**: you get a sample $\mathbb{P}'(X)$,
2. **Transfer Learning**: you get another sample $\mathbb{P}'(X, Y)$,
3. You simply **get more outputs** $\mathbb{P}'(Y)$.

How can you get a better model of the true $\mathbb{P}(X, Y)$? **A Causal Graph can help**, indeed, it may be the only way to rigorously formalize how to combine them [@scholkopf2012causal].

Dealing With Unknown Bias: From Post-Stratification To Optimal Transportation

Biases Of Nowcasting i

Most often, datasets are collected for a specific purpose. For instance

- web traffic to optimize the navigation of web sites
- credit card tickets to target financial services
- suppliers and clients as declared to regulators
- etc.

In all these examples the dataset is (almost) an i.i.d. sample of the distribution of interest, because data are collected close to the populaion of interest: users of web sites, users / usages of credit cards, companies under a specific regulation, etc.

But to implement nowcasting, your distribution is made of all the economic or physical entities of this kind: all web sites, all consumers spending in shops, all companies, etc.

As a consequence these dataset have **a collection bias**: part of the population of interest is not considered at all during the collection process (only users using internet to get awareness on a company, only consumers having credit specific use of their cards, a specific type of companies, etc).

Biases Of Nowcasting ii

But it is not all: for systematic investment you need to connect (to *map*) entities in your dataset to tradable instruments. Here again there is another bias, a **blind spot bias**. You will only have traffic on companies using a specific web technology, credit card expenses will not make sense for all companies,

To handle these biases, you need to be creative during your investigation, going back and forth in **testing different comparisons with reference datasets**.

Once it is done, you can

- estimate the biases and propose a correction, that is the porpose of **post-stratification**,
- you do it an interactive way, that **introduces an exploration-exploitation process**

Dealing With Unknown Bias: From Post-Stratification To Optimal Transportation

Post-Stratification

Post-stratification i

> “Broadly speaking, post-stratification refers to any method of data analysis which involves forming units into homogeneous groups after the sample has been taken.” [Zhang, 2000]

The way it is usually expressed in the literature (Survey Theory) is

- You start with **strata** $s \in \mathcal{S}$ that are disjoint subsets created from a categorical variable (a state, an industry, the age, etc),
- You want to apply weights $(w_s)_s$ to these strata such that the weighted average of an observations $(x_s)_s$ is as close as possible to the desired expected value $\bar{X}$: $\sum_s w_s x_s \simeq \bar{X}$. Keep in mind that $\bar{X}$ and each x_s can be a vector is you want to control simultaneously for several biases.
- But you want the weights w_s to be as close as $a := 1/\#\mathcal{S}$ as possible; you express this using a distance function $\sum_s aG(w_i/a)$ (see [Deville et al., 1993]).
- Hence you end up with a constrained optimization that, when $G(r) := (r-1)^2/2$ boils down to

$$w_i = a \cdot \left\{ 1 + (\bar{X} - \sum_s a x_s) \left(\sum a x_s^T x_s \right)^{-1} x_i^T \right\}.$$

Post-stratification ii

A lot of variations have been proposed. In essence this approach allows,

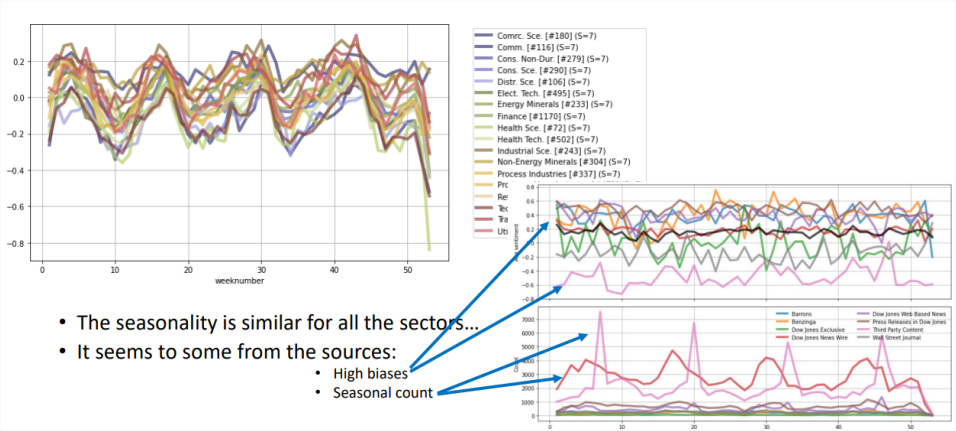
- once you identified groups of observations on which you have an external reference,
- to adjust weights on your observed sample to get them as close as possible to this external reference.

👉 It can probably be reformulated as **optimal transportation problem**...

I am not sure of the added value, and in any case

- uncertainty due to the size of your sample has to be taken into account,
- it is “easy” as long as you restrict youself to groups of observations.

It Is The Same For Seasonalities (example of a “NLP sentiment”)



- The seasonality is similar for all the sectors...
- It seems to some from the sources:
  - High biases
  - Seasonal count

Note that not desired seasonalities can be considered as biases and processed the same way (NLP have other biases [Li and Lehalle, 2021]).

Dealing With Unknown Bias: From Post-Stratification To Optimal Transportation

Covariate Shift

If A Bias Was Nothing Else Than A Covariate Shift?

A **covariate shift** happens [Sugiyama and Kawanabe, 2012] when

- **you trained a model on a distribution $\mathbb{P}_{\text{tr}}$**
- **you have to use it on another distribution $\mathbb{P}_{\text{te}}$.**

Ideally you want to learn the change of measure from $\mathbb{P}_{\text{tr}}$ to $\mathbb{P}_{\text{te}}$. Indeed, give an loss function $\ell(x, y, \theta)$, where θ are the parameters of a lodel, you can formally write

$$\mathbb{E}_{(x,y)\sim\mathbb{P}_{\text{te}}}(\ell) = \mathbb{E}_{(x,y)\sim\mathbb{P}_{\text{tr}}}\left(\frac{\mathbb{P}_{\text{te}}}{\mathbb{P}_{\text{tr}}} \cdot \ell\right). \tag{1}$$

You face the usual diificulties:

- what if the support of $\mathbb{P}_{\text{te}}$ is not included in the one of $\mathbb{P}_{\text{tr}}$?
- how to discretize? [Birgé and Rozenholc, 2002]

👉 it is not very different of the bias correction problem (but now it is its continuous version).

Kernel Methods For Covariate Shift i

Following [Gretton et al., 2009], if you focus on the effect of the expectation operator $\mu(\mathbb{P}) := \mathbb{E}_{x\sim\mathbb{P}}(\Phi(x))$, restricting yourself to mappings Φ into a feature space that is a Reproducing Kernel Hilbert Space (i.e. it can naturally be represented by a kernel $k(x,y) := \langle \Phi(x), \Phi(y) \rangle$),

then the weight function $\beta(x)$ solution of

$$\begin{array}{ll} \textbf{min} & \|\mu(\mathbb{P}_{te}) - \mathbb{E}_{x\sim\mathbb{P}_{tr}}(\beta(x)\Phi(x))\| \\ \text{s.t.} & \beta(x) \geq 0, \\ & \mathbb{E}_{x\sim\mathbb{P}_{tr}}(\beta(x)) = 1 \end{array} \tag{2}$$

is the desired one.

I.e. when $\mathbb{P}_{te}$ is absolutely continuous with respect to $\mathbb{P}_{tr}$, then $\mathbb{P}_{te}(x) = \beta(x)\mathbb{P}_{tr}(x)$.

Moreover, this problem is **convex in** β.

Note that it is again a transportation problem.

Kernel Methods For Covariate Shift ii

Moreover, in practice, setting

$$K_{i,j} := k(x_i^{\text{tr}}, x_i^{\text{tr}}) = \langle \Phi(x_i^{\text{tr}}), \Phi(x_j^{\text{tr}}) \rangle, \quad \kappa_i := \frac{N_{\text{tr}}}{N_{\text{te}}} \sum_j k(x_i^{\text{tr}}, x_j^{\text{te}}) = \frac{N_{\text{tr}}}{N_{\text{te}}} \sum_j \langle \Phi(x_i^{\text{tr}}), \Phi(x_j^{\text{te}}) \rangle;$$

one can write

$$\left\| \frac{1}{N_{\text{tr}}} \sum_{i=1}^{\text{tr}} \beta_i \Phi(x_i^{\text{tr}}) - \frac{1}{N_{\text{te}}} \sum_{i=1}^{\text{te}} \Phi(x_i^{\text{te}}) \right\| = \frac{1}{N_{\text{tr}}^2} \beta^T K \beta - \frac{2}{N_{\text{tr}}^2} \kappa^T \beta + c.$$

This allows, once a kernel is chosen, to find the weight function β as a solution of this problem (for a well chosen ϵ that is of the order of the standard deviation of teh empirical average of β)

$$\begin{array}{ll} \textbf{min} & \frac{1}{2} \beta^T K \beta - \kappa^T \beta \\ \text{s.t.} & 0 \geq \beta(x) \geq B, \\ & |\sum_i \beta_i - N_{\text{tr}}| \leq N_{\text{tr}} \epsilon \end{array} .$$

☞ Any similar approach is welcome, but notice that this one is not very concerned by a discretization problem (indeed ϵ is taking care of this).

The Power Of Causality Indentification

Positioning: Datasets? Specific Difficulties In Finance?

Dealing With Unknown Bias: From Post-Stratification To Optimal Transportation

The Power Of Causality Indentification

The Power Of Causality Indentification

Recent Results in Causality Identification

DAG For A Joined Probability Distribution

The goal of Causal Inference is to find "*simple expressions*" for $\mathbb{P}(X_1, \ldots, X_d)$. *Simplicty* is not a well defined concept, here we talk about **splitting the joined distribution in independent blocks**. (Peters, Janzing, and Schölkopf 2017) proposes to focus on the **causal (or disentangled) factorization**:

$$\mathbb{P}(x_1, \ldots, x_d) = \prod_{j=1}^{d} \mathbb{P}(x_j | \mathrm{PA}_j),$$

following the seminal formal work done by (Pearl 2009) and (Spirtes et al. 2000).

Moreover, they propose to use **Structural Causal Models** as generative model (showing that their can represent all reasonnable distributions), allowing to write

$$X_j := f_j(\mathrm{PA}_j, N_j);\ i \neq j \Rightarrow N_i \perp\!\!\!\perp N_j,$$

where PA_j are the *parents* of X_j in a Directed Acyclic Graph (DAG). The simplest example being: $X = N_X$, $Y = f(X) + N_Y$, where $N_X \perp\!\!\!\perp N_Y$.

SCM: Structural Causal Models = Adjacency Matrix + Functions

A simple parametric model for a Structural Causal Model is

1. An adjacency matrix B_{ij} connecting Effect X_i to its parents X_j when $B_{ij} = 1$;
2. For each Effect X_i, a function $f_i(\mathrm{PA}_i)$ such that $X_i = f_i(\mathrm{PA}_i) + N_i$.

In this formulation, we have few nice immediate properties:

- $(X_i - f_i(\mathrm{PA}_i)) \perp\!\!\!\perp \mathrm{PA}_i$. This will be useful to check that the correct parents of X_i have been found.
- One can test *interventions* on the model: setting N_i to zero or $f_i(\cdot) = 1$ for instance.
- It is also possible to understand what are the consequences of having access to another sample of PA_i (i.e. Semi-Supervised Learning), or to another sample (X_i, PA_i) (i.e. Transfer Learning).

Correspondence Between Causality Inversion And Gaussianity

In the 50ties, the **Darmois-Skitovic theorem** made an equivalence between invertision of causality and Gaussianity the following way (Theorem 4.3 of (Peters, Janzing, and Schölkopf 2017)):

> *Let $X_1, \ldots, X_d$ be independent, non-degenerate random variables. If there exist non-vanishing coefficients $a_1, \ldots, a_d$ and $b_* 1, \ldots, b_d$ (that is, for all $i: a_i \neq b_i$ such that the two linear combinations: $\ell_1 = \sum_i a_i X_i$ and $\ell_2 = \sum_i b_i X_i$ are independent, then each X_i is normally distributed.*

The **Independent Component Analysis** exploits also the non-invertibility of non-Gaussian sources. It does exactly the reverse of the Darmois-Skitovic theorem; quoting (Oja, Erkki and Hyvarinen, A. 2000):

> *The Central Limit Theorem, a classical result in probability theory, tells that the distribution of a sum of independent random variables tends toward a gaussian distribution, under certain conditions. Thus, a sum of two independent random variables usually has a distribution that is closer to gaussian than any of the two original random variables. [...] Therefore, we could take as w a vector that maximizes the nongaussianity of $w^T x$ [to recover the independent original sources].*

The `FastICA` algorithm maximizes approximations of the neg-entropy (i.e. the entropy of a Gaussian approximation of a random variable minus its empirical entropy) to recover independent non-Gaussian sources that are mixed a linear way.

Hand Waving The Characterisation Of Anticausality i

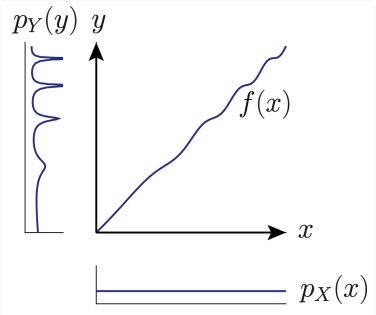


Figure 1: A natural deep learning model for SCM

Before being technical, have a look at the simplest model for a Cause to Effect without any noise: $C = N_C$ and $E = f(C)$. It is interesting to notice that, **provided that $f(\cdot)$ and the cause N_C are "independent enough", the distribution of the effect E is concentrated around regions where $f(\cdot)$ is "flat"**.

Hand Waving The Characterisation Of Anticausality ii

To be more accurate, following (Daniusis et al. 2012), it is useful to define "independent enough" as

$$\mathbb{C}\mathrm{ov}(\log f', p_C) = \int_0^1 \log f'(x)\, p_C(x)\, dx - \int_0^1 \log f'(x)\, dx = 0,$$

This is equivalent to $\int_0^1 \log f'(x)\, p_C(x)\, dx = \int_0^1 \log f'(x)\, dx$.

Using the notation $g := f^{-1}$, the upper equality reads also $\int_0^1 \log g'(y)\, p_C(y)\, dy = \int_0^1 \log g'(y)\, g'(y)\, dy$, and hence

$$\mathbb{C}\mathrm{ov}(\log g', p_E) = \int (g'(y) - 1) \log g'(y)\, dy = D(g' \| v) + D(v \| g'),$$

where v is a uniform distribution on $[0, 1]$ And $D(\cdot\|\cdot)$ is the relative entropy. As a consequence $D(g'\|v) + D(v\|g') \geq 0$. This formalizes the fact that "**when f' and the cause are independent, the derivative of the invert of f and the effect are not**". This is one more not invertibility property of cause and effect seen from an SCM perspective.

A Clean Characterization Of Non Invertibility Of Causality i

Following (Hoyer et al. 2008), it is starightforward to **characterize the invertibility of the relation between cause x and effect y by the existence of function solving an ODE**:

- set $y = f(x) + n$ and $x = g(y) + \tilde{n}$,
- we will derive a joined differential equation for f, $\nu = \log p_n$ and $\xi = \log p_x$ when the causality is invertible by differentiating two expressions of $\pi(x,y) := \log p(x,y)$,
- one corresponds to the causal expression of (x,y): $\log p(x,y) = \log p_n(y - f(x)) + \log p_x(x)$ and the other to the anticausal one: $\log p(x,y) = \log p_{\tilde{n}}(x - g(y)) + \log p_y(y)$.
- that for we need a notation for $\tilde{\nu} = \log p_{\tilde{n}}$ and $\eta = \log p_y$.

A Clean Characterization Of Non Invertibility Of Causality ii

The goal is to make the ration between two partial derivatives of the anticausal version of π to cancel any influence of x:

$$\frac{\frac{\partial^2 \pi}{\partial x^2}}{\frac{\partial^2 \pi}{\partial x \partial y}} = \frac{\tilde{\nu}''(x - g(y))}{-\tilde{\nu}''(x - g(y))g'(y)} = \frac{1}{g(y)}.$$

It implies that, provided the anticausal expression is valid:

$$\partial(\frac{\partial^2 \pi}{\partial x^2} / \frac{\partial^2 \pi}{\partial x \partial y})/\partial x = 0.$$

And now if doing the same on the causal version of π, one reads:

$$\frac{\frac{\partial^2 \pi}{\partial x^2}}{\frac{\partial^2 \pi}{\partial x \partial y}} = \frac{\nu''(f')^2 - \nu' f'' + \xi''}{-\nu''(y - f(x))f'(x)}.$$

A Clean Characterization Of Non Invertibility Of Causality iii

The partial derivative of this expression with respect to x is a little heavy; it boild down to

$$\partial \left(\frac{\frac{\partial^2 \pi}{\partial x^2}}{\frac{\partial^2 \pi}{\partial x \partial y}} \right) / \partial x = -2f'' + \frac{\nu' f'''}{\nu'' f'} - \frac{\xi'''}{\nu'' f'} + \frac{\nu' \nu''' f'}{(\nu'')^2} - \frac{\nu' (f'')^2}{\nu'' (f')^2} - \xi'' \frac{\nu'''}{(\nu'')^2} + \xi'' \frac{f''}{\nu'' (f')^2}. \tag{3}$$

That has to be zero if the causal and anticausal expressions for $p(x, y)$ both hold.

A Clean Characterization Of Non Invertibility Of Causality iv

Before simplifying this ODE, have a look at it when the noises are Gaussians, i.e. $\nu(x) = c - x^2/2, \nu'(x) = -x, \nu'' = -1$ and $\nu''' = 0$. The same for ξ. It is immediate that one should have

$$-2f'' + x\frac{f'''}{f'} - x\frac{(f'')^2}{(f')^2} + \frac{f''}{(f')^2} = 0.$$

This entirely describes the set of causal relations that are invertible in a Gaussian setup. **Of course linear functions are part of this set, but not only.** Indeed some intrications between the causal function f and the noises are possible that allows causal inversion; see Theorem 1 of (Hoyer et al. 2008).

Coming back to the generic ODE (3), we can read it as

$$\xi'''(x) = \xi''(x)\, G(x,y) + H(x,y), \quad G := -\frac{\nu'''f'}{\nu''} + \frac{f''}{f'}, \quad H := -2\nu''f''f' + \nu'f''' + \frac{\nu'\nu'''f''f'}{\nu''} - \frac{\nu'(f'')^2}{f'}.$$

That can be solved in $z := \xi''$ that is specified by $z(x_0)$ as soon as $\nu''(y - f(x))f'(x) \neq 0$.

This means that

- once a relation $y = f(x) + n$ exists (keep in mind that $\nu = \log p_n$)
- if you find $\xi = \log p_x$, the solution of the upper equation,
- then it specifies the distributions of x **such that the anticausal relation** $x = g(y) + \tilde{n}$ **is true too**.

(you can play with $f(x) = -x$ and Gumbel distributions for x and n if you want an example)

The Power Of Causality Indentification

In Practice: Have A Look At Simple Inversions (i.e. Basic Anticausal Relations)

In Practice: Have A Look At Simple Inversions i

A notebook available at (Lehalle 2022) simulate two very simple examples: a discrete simple SCM and a continuous one. These figures show how $E - \hat{f}(C)$ is more independent of the cause C than $C - \hat{g}(E)$ is independent of E.

The continuous model is simply $X = N_X, Y = 1 + X + X^2 + 1.5N_Y$.

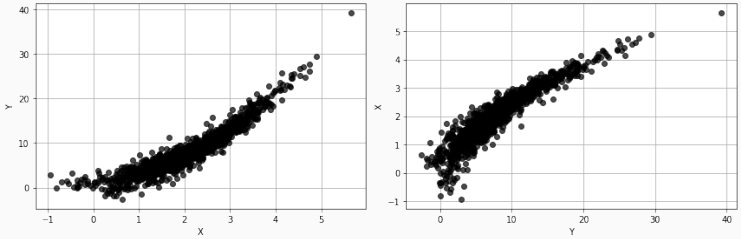


Figure 2: The simple "linear" additive model

In Practice: Have A Look At Simple Inversions ii

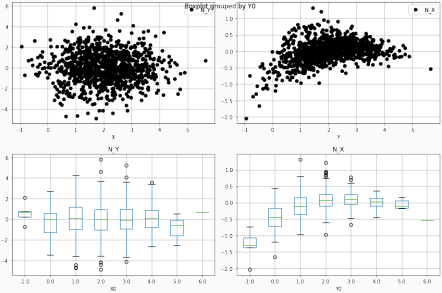


Figure 3: Residuals of linear regression on the continuous model

One sample of 1,000 realizations of the model, seen from a $Y = f(X)$ (left panel) or from a $Y = g(X)$ perspective (right panel).

On the Top: The scatter plots of the regression of Y by X (left Panel) or of X by Y (right Panel), on the bottom: The conditioning by values of the potential "cause" of the effect; the right distributions seem less i.i.d. than the left ones.

This effect is very interesting from a statistical perspective since it means that the correct (i.e. causal) representation recovers i.i.d. or stationarity properties that financial datasets lack. It should also provide ways to be robust to covariate shift.

In Practice: Have A Look At Simple Inversions iii

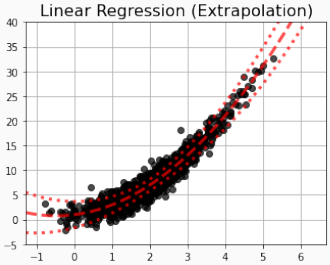
In a linear setup, it is worthwhile to note that the asymmetry between a regression of Y by X or of X by Y is obvious (because the corresponding projections are orthogonal). From a generalization perspective keep in mind that confidence intervals of a linear regression are larger (philosophically corresponding to an extrapollation):

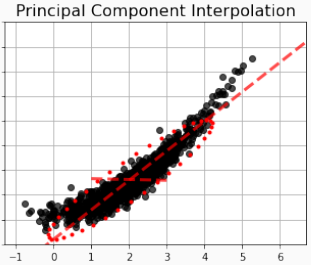
$$y(x) \in \left[\hat{y} \pm \tau_{n-2}\, \sigma_y \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{(n-1)\, \sigma_x^2}} \right].$$

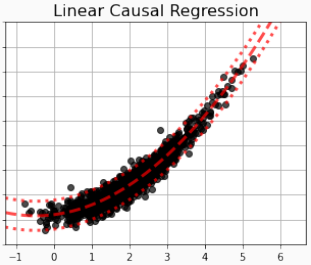
With causality you expect it to be independent from the distance to $\bar{x}$, the empirical middle of the sample of n occurrences of x, since the prediction $Y(X) \perp\!\!\!\perp X$.

It is how it looks in reality

Same dataset: three models and their **Prediction Intervals**







The Power Of Causality Indentification

Do You Want More x Or More y?

Theoretical Advantages: Interpolation, Extrapolation, and Counterfactual Questions

As stated earlier, natural datasets have (domain) shifts or unknown interventions, especially when they are collected from different sources. Seen **from a statistical viewpoint they exhibit biases**, that stems from the fact that they are collected from (slightly) different distributions $\mathbb{P}'(C), \mathbb{P}'(E)$ or $\mathbb{P}'(E, C)$ than the *ground true ones*.

Understanding the causal graph can help to patch these distributions together to obtain a better estimate of the true $\mathbb{P}(E, C)$ (see (Schölkopf et al. 2012)). For instance **having more data on the effects is more useful than having more data on the causes**, simply because observations of the effects contains information on the joined distribution of the causes and effects, while observations of the cause contains no information on this joined distribution.

Moreover, **without a causal representation of a dataset, a learnt relation is only accurate within identical experimental conditions**.

Focus: From Semi-Supervised Learning To Unsupervised Learning In A Causal Setup

Another advantage of a Structural Causal Model is that it tells what you do when you obtain more data. Here we focus on the supplementary information of occurences of the cause C but not of the effect, that is the **Semi-Supervised Learning (SSL) case**. Usually, under regularity conditions, one hopes that *when a new cause is close to a known cause, then both effects are close*.

A typical case for that is when one operate a clustering on variables X, say using a **mixture of K Gaussians** or a **Self-Organizing Map of K prototypes** (leading to a Voronoï split of the space around these prototypes, see (Kohonen 2001)), and then associate a label Y to each observation. Idealy they are K different labels, corresponding to a one-to-one mapping between classes and clusters.

If the label Y is the effect and the X contains the causes plus noise, an ideal (additive) model is: $X = C + N_X$, $C = N_C, Y = f(C) + N_Y$. It is indeed clear that **a non supervised clustering is efficient if the variables to be clustered contained all the information on the labels**, i.e. if they contain information on the effects more than on the cause. The trivial case is of course a clustering on labels…

The SCM tells us that $Y - f(C)$ is independent of the cause C, and hence more information on the cause will not help to understand the labels / effect Y. The natural smoothness assumption of SSL is a way to link information on the cause and information on the effect. **It is of course linked with the needed correlation of $(f^{-1})'$ and p_E previously seen**.

Time To Conclude

Positioning: Datasets? Specific Difficulties In Finance?

Dealing With Unknown Bias: From Post-Stratification To Optimal Transportation

- Post-Stratification
- Covariate Shift

The Power Of Causality Indentification

- Recent Results in Causality Identification
- In Practice: Have A Look At Simple Inversions (i.e. Basic Anticausal Relations)
- Do You Want More x Or More y?

Data Curation Via Causality Detection

As explained before, Data Curation for financial markets goes further than data integrity. It needs to perform what could be called **Active Post-Stratification**, at the middle of **Active Learning** and standard **Post-Stratification**. The goal is to identify biases usually coming from

- the collection bias (on the **inputs**) that is often directed towards *mainstream economic entities* (users of technology, large companies, large cities, etc),
- *blindspots on some outputs* of these data (activity of non-retail facing companies will not be caused by credit card tickets, governmental entities are often exempted from reporting, etc),
- *covariate shifts* (changes of economic context, or in the collection methodology, leading to *unknown interventions*),
- *heterogeneity of data sources* (News articles come from mainstream press and local newspapers, as from governmental agencies).

We have seen that bias identification (**post-stratification**) is close to **optimal transportation** and that recent progresses in causality identification could help to design **strategies to combine different instances of the same phenomena** (cf. (Schölkopf et al. 2012)).

A Lot Of Open Questions

- How the link between post-stratification and optimal transportation can help?
- What is the effect of interative explorations during post-stratification?
- There is a link with active learning)
- When to stop post-stratification?
- Causality identification seems to boil down to good test of independence, what are the proper ones?
- How to avoid spurious causality detection?
- What is the link between causality and control?

Any question?

charles.lehalle@adia.ae

Nothing is new: The day to day work of a data scientist was already well defined in the early XXth century (1930). In *The Man without Qualities*, Robert Musil writes:

> *[...] as happens so often in life, you [...] find yourself facing a phenomenon about which you can't quite tell whether it is a law or pure chance; that's where things acquire a human interest. Then you translate a series of observations into a series of figures, which you divide into categories to see which numbers lie between this value and that, and the next, and so on [...]. You then calculate the degree of aberration, the mean deviation, the degree of deviation from some arbitrary value [...] the average value [...] and so forth, and with the help of all these concepts you study your given phenomenon.*

References i

Ankan, A., Wortel, I. M., and Textor, J. (2021).
Testing graphical causal models using the r package "dagitty".
Current Protocols, 1(2):e45.

Bengio, Y., Deleu, T., Rahaman, N., Ke, R., Lachapelle, S., Bilaniuk, O., Goyal, A., and Pal, C. (2019).
A meta-transfer objective for learning to disentangle causal mechanisms.
arXiv preprint arXiv:1901.10912.

Birgé, L. and Rozenholc, Y. (2002).
How many bins must be put in a regular histogram.
Preprint du LPMA, 721.

Birgé, L. and Rozenholc, Y. (2006).
How many bins should be put in a regular histogram.
ESAIM: Probability and Statistics, 10:24–45.

References ii

Briere, M., Lehalle, C.-A., Nefedova, T., and Raboun, A. (2020).
Modeling transaction costs when trades may be crowded: A bayesian network using partially observable orders imbalance.
Machine Learning for Asset Management: New Developments and Financial Applications, pages 387–430.

Cardaliaguet, P. and Lehalle, C.-A. (2018).
Mean field game of controls and an application to trade crowding.
Mathematics and Financial Economics, 12(3):335–363.

Chen, L., Pelger, M., and Zhu, J. (2019).
Deep learning in asset pricing.
arXiv preprint arXiv:1904.00745.

Daniusis, P., Janzing, D., Mooij, J., Zscheischler, J., Steudel, B., Zhang, K., and Schölkopf, B. (2012).
Inferring deterministic causal relations.
arXiv preprint arXiv:1203.3475.

References iii

Deville, J.-C., Särndal, C.-E., and Sautory, O. (1993).
Generalized raking procedures in survey sampling.
Journal of the American statistical Association, 88(423):1013–1020.

Glymour, C., Zhang, K., and Spirtes, P. (2019).
Review of causal discovery methods based on graphical models.
Frontiers in genetics, 10:524.

Gretton, A., Fukumizu, K., Teo, C., Song, L., Schölkopf, B., and Smola, A. (2007).
A kernel statistical test of independence.
Advances in neural information processing systems, 20.

Gretton, A. and Györfi, L. (2010).
Consistent nonparametric tests of independence.
The Journal of Machine Learning Research, 11:1391–1423.

Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., and Schölkopf, B. (2009).
Covariate shift by kernel mean matching.
Dataset shift in machine learning, 3(4):5.

References iv

Guo, L. and Modarres, R. (2020).
Nonparametric tests of independence based on interpoint distances.
Journal of Nonparametric Statistics, 32(1):225–245.

Hoyer, P., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. (2008).
Nonlinear causal discovery with additive noise models.
Advances in neural information processing systems, 21.

Kalainathan, D., Goudet, O., and Dutta, R. (2020).
Causal discovery toolbox: Uncovering causal relationships in python.
J. Mach. Learn. Res., 21:37–1.

Ke, N. R., Bilaniuk, O., Goyal, A., Bauer, S., Larochelle, H., Schölkopf, B., Mozer, M. C., Pal, C., and Bengio, Y. (2019).
Learning neural causal models from unknown interventions.
arXiv preprint arXiv:1910.01075.

Kohonen, T. (2001).
Learning vector quantization.
In *Self-organizing maps*, pages 245–261. Springer.

Lehalle, C.-A. (2022).
Anticausal inference on sample data using additive noise models.
`https://colab.research.google.com/drive/1usn5PUDnEvvLwOFjbO2eJcUhiWO8pAI9?usp=sharing`.

Li, M. and Lehalle, C.-A. (2021).
Do word embeddings really understand loughran-mcdonald's polarities?
arXiv preprint arXiv:2103.09813.

Lin, J.-J., Chang, C.-H., and Pal, N. (2015).
A revisit to contingency table and tests of independence: bootstrap is preferred to chi-square approximations as well as fisher's exact test.
Journal of biopharmaceutical statistics, 25(3):438–458.

References vi

Oja, E. and Hyvarinen, A. (2000).
Independent component analysis: algorithms and applications.
Neural networks, 13(4-5):411–430.

Pearl, J. (2009).
***Causality*.**
Cambridge university press.

Peters, J., Janzing, D., and Schölkopf, B. (2017).
***Elements of causal inference: foundations and learning algorithms*.**
The MIT Press.

Ratkovic, M. (2014).
Balancing within the margin: Causal effect estimation with support vector machines.
Department of Politics, Princeton University, Princeton, NJ.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. (2012).
On causal and anticausal learning.
arXiv preprint arXiv:1206.6471.

Scutari, M. (2021).
Introduction to bayesian networks: How we can use them as probabilistic and causal models.
https://www.bnlearn.com/about/slides/slides-zhaw21.pdf.
Accessed: 2020-11-06.

Spirtes, P., Glymour, C. N., Scheines, R., and Heckerman, D. (2000).
Causation, prediction, and search.
MIT press.

Sugiyama, M. and Kawanabe, M. (2012).
Machine learning in non-stationary environments: Introduction to covariate shift adaptation.
MIT press.

Tian, J. and Pearl, J. (2001).
Causal discovery from changes.
In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 512–521.

Zhang, L.-C. (2000).
Post-stratification and calibrationa synthesis.
The American Statistician, 54(3):178–184.