How Confrontation Of Information And Flows Shapes Prices

From Data To Decision (And The Related Mathematical Tools)

Charles-Albert Lehalle

Professor, CMAP, Ecole Polytechnique, IP-Paris

(Louis Bachelier Fellow, Paris, France)

CERMICS, Les Ponts

May 21, 2025

Goal For Today

The way prices form on financial markets can appear to be very simple: **a double auction game takes place** in real-time, where buyers meet sellers.

This apparent simplicity hides the fact that decisions and actions on financial instruments are, in general, synchronised: everyone wants to **buy Pfizer when it discovers the vaccine**, or buy NVIDIA when is in apparently useful for ML.

And this is natural: when information is disclosed, market participants deduce decisions, driving the way contribute to the auction game (buy or sell).

They are a few caveats:

1. If extracting decisions from information is costly, how do one makes money to get paid for this effort? Participants doing their homework should be able to take positions before the others: this is the **Grossman-Stiglitz Paradox** .
2. When the information becomes easy to understand: everyone rushes and (since nobody knows exactly where the price should be) the demand/offer imbalance makes an overshoot.
3. Then it comes back to a more reasonable price (but in between new information flows).

The Information - Decision Lifecycle

In short, the process is the following

1. Some participants harvest information the way they can, the easiest to get, the less revenue they will make.
   ☞ With the **recent availability of alternative data** (Satellite images, credit cards, geolocation, texts, etc): the menu is large.
2. They **take decisions anticipating the effect of their pressure**: if you discover an information α, leading to an improvement of the price $\Delta S(\alpha)$, and if the effect of your pressure is square-root (via Kyle's λ, [Kyle, 1985]), your expected profit for trading $Q(\alpha)$ is
$$\max_{Q} \mathbb{E}\left[Q(\alpha)(\Delta S(\alpha) - \lambda\sqrt{Q(\alpha)})\middle|\alpha\right] \Rightarrow Q^*(\alpha) = \left(\frac{2\Delta S(\alpha)}{3\lambda}\right)^2.$$
3. This is a little more complicated because there is a **mean field effect**: the λ is a function of what the others are doing (being alone understanding News makes it small).
4. This is even more difficult because they need to disantangle the move due to their impact from **the natural spread of the information** (i.e. tomorrow more participants will have understood).

This cycle takes place at different time scales, simultaneously on different financial instruments (that are nevertheless related by their economical underlying, like two companies from the same sector), at different time-horizons and

☞ **this is the way capital is allocated to projects**.

Outline

The Grossman-Stiglitz Paradox

The Informational Content Of Alternative Data

The Process Of Taking Decisions

Moving (Or Forming?) The Price

Optimal Control Of The Trading Process

Next Section – The Grossman-Stiglitz Paradox

The Grossman-Stiglitz Paradox

This paradox came to counter the CAPM idea that at any time, prices reflect the amount of available information on projects [Grossman and Stiglitz, 1980].

What does it means in practice?

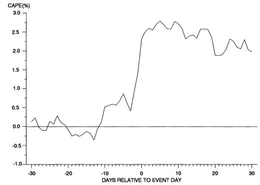


FIG. 1.—Cumulative average prediction errors for 79 portfolios of buy recommendations.

20 years ago, analysts recommendations took days to be in the prices [Beneish, 1991], 10 years ago, it was done in one year, and now it is less than a hour:

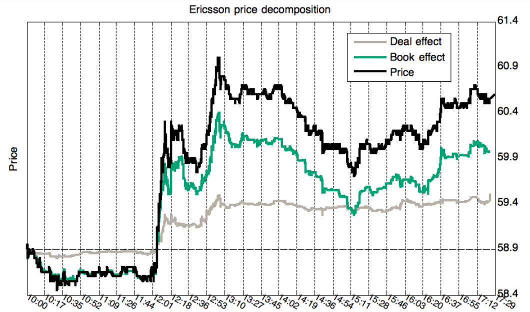
- long ago, you had to be present at the conference (with an audio system to listen in rooms around),
- then brokers broadcasted their morning meetings and sent pdf files,
- intermediaries (lke Bloomberg or FactSet) provide the information in a real-time feed and in a machine readable format,
- Natural Language Processing in general, and more recently Language Models, made this information very easy to get.

This process is highly regulated.

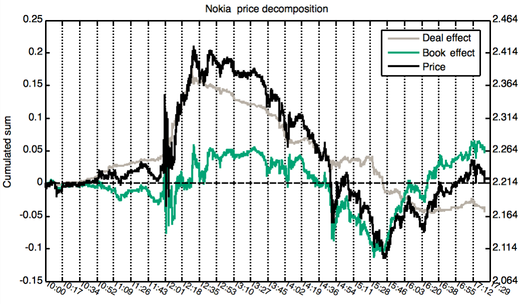
The Information Is Not Always Easy To Understand?

In [Besson and Lehalle, 2014], we try to split price moves caused by trades or just “quote shifts”, we provided aggregated statistics but here are two anecdotical cases, the 18th of Oct. 2012; it shows how news impact the price, it somehow shows participants have several ways to impulse permanent impact:

easy to process news



difficult to process news



Next Section – The Informational Content Of Alternative Data

The Informational Content Of Alternative Data

With the raise of **data sciences** (i.e. big data + machine learning), the availability of new datasets (and the means to draw conclusions from data) dramatically evolved the last 10 years. Some examples

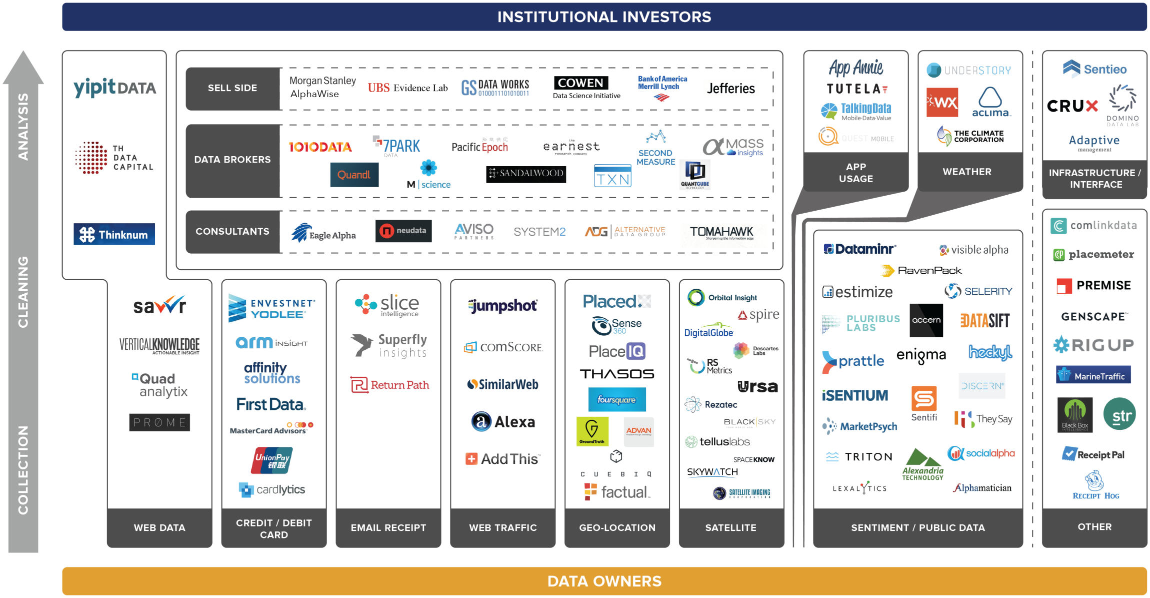
- **Supply chain**: extracted from reports, observed in the inventories of plants and warehouses, follwing ships and trucks, etc.
- **Geolocation** thanks to the Internet of Things: mobile phones, cars and trucks, wifi access points, etc
- **Satellite images**: different wavelengths and sensors for light, heat, biomass, weather, humidity, etc on the one hand, or images for different activities in car parks, warehouses, etc
- **Texts**: from financial texts like transcripts of Earning Calls and Financial News to Patents, job postings, comments on market places and social networks, etc.
- etc.

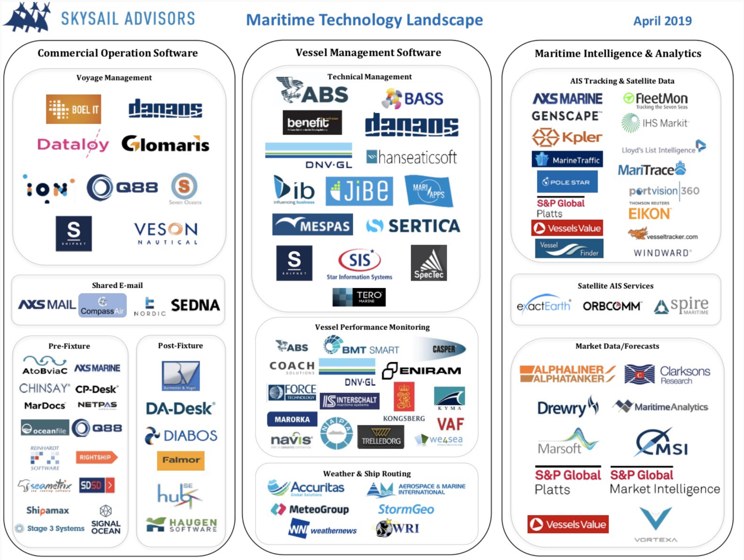
At first sight, it impacts the way information diffuses into prices in two ways

- a **speed advantage**: exploiting information an automated way beats human capabilities ("blink" of an eye is 500ms long...)
- a **cross sectional advantage**: browse quickly information to compare the value of assets is now possible at scale (read 1,000 documents with the same methodology in less than 2 minutes?)

ALTERNATIVE DATA STACK

alternativedata.org





And you can zoom (here data on maritime traffic and shipping only. This means that you have access to microscopic scenario, that you can decide to aggregate (or not), to compete with official agencies' numbers.

Of course it is not that simple, you do not observe 100% of the economic inputs, storages and outputs. Neither your observations have the same time scale (mix of real-time, weekly, etc).

Different Viewpoints Give Birth To Different Biases

Alternative data are usually generated for a **natural use ourside of the financial world**:

- Web traffic data are an iid sample of the users of one web site (used to optimize the traffic on this web site);
- Credit card data are an iid sample of customers of a bank to be targeted by advertisment for products;
- The tweets about a brand are reflecting what customers think about the products;
- etc.

But when you want to **nowcast** [BańBura et al., 2014], you want to build a realistic picture of the economic world:

- Web traffic data **are not** an iid sample of existing retail web sites;
- Credit card data **are not** an iid sample of people spending money in stores;
- The tweets about products **are not** reflecting the relative interest of customers for brands;
- etc.

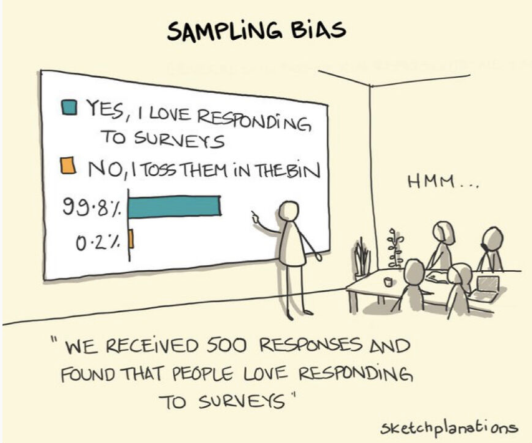
Choose Your Viewpoint: You have Chosen Your Bias

To nowcast the state of the economic entities that are the underlyings of financial products, you need to

- collect as much data as possible,
- patch the datasets you have,
- inderstand (and correct) their biases.

You face two kinds of biases:

- The Collection Bias,
- The Targeting Bias.



Understand What You Are Observing: (1) The Collection Bias

Each sensor has its own **bias due to the collection process**:

- **Credit Cards** only observes consumers having credit cards, or using money management Apps,
- Most equipments of **satellites** only gives good pictures when there is no cloud,
- During **Earning Calls**, CFO and CIO of companies only speak of positive points of their companies,
- Etc

You need to identify these biases and try to correct them:

- For demographics, you can access to public agencies statistics and correct (i.e. re-weight) for the identified biases
- But usually it is a dataset per dataset reasoning (two credit card datasets can have different biases).

☞ These biases should be identify as close as possible to the data collection, it cannot be done without **a field expertise in domain documented by the dataset**.

Understant What You Are Observing: (2) The Targeting Bias

An important step is the **matching of the observed entities with investible instruments**:

- **Satellite images** and **geolocation**: a polygon of Lat x Lon gives you an Area of Interest, that you have to **match** with an economic activity shares by economic entities (like “this is a field of corn” –for Future contracts– or “this is a Starbucks” –for Listed Equities–)
- **Credit cards**: you need to **match** shop names and brands to companies and companies to Listed Equities,
- **Financial News**: you need to **match** a paragraph of a News to a company (i.e. a Listed Equity) or an economic entity (i.e. Inflation),
- Etc

This matching introduces biases: you can **systematically miss**

- Companies who are not owning their inventories but lending them
- A misspelled brand name, or miss-match a brand that has been recently sold to another company
- Etc

☞ And all this matching has to be **Point in Time** to be able to replay the past without any **survivorship bias**.

Next Subsection – Contexts, Biases, Post-Stratification And Covariate Shift

What Kind Of Biases Are We Talking About?

A typical example is a population of geolocalized mobile phones in US malls, you may want to compare

- the demographics of this population to the one of the US population,
- the brands presents in the malls of the dataset with the retail brands in the US,
- the usual overshoot of consumption during the Back Friday with the geolocation.

It is a question of **choosing a reference model**. For instance which population do you have in mind

- the whole US population? US consumers?
- US consumers in malls? US consumers of brands that are in malls?
- US consumers of brands that are in your dataset?

You have to choose a reference that is not too far away from your sample, to **reflect the natural informational content of the dataset**.

The ideal reference is related to what you want to do with the data (do you really want to use the data very far away from their natural distribution?).

Post-stratification

> “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):

- 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 are vectors (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 $\sum_s a \cdot G(w_i/a)$ (see [Deville et al., 1993]).
- Hence you end up with a constrained optimization that, when $G(r) := (r-1)^2/2$ boils down to

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

Impact Of Bias Correction

The correction w_i is not certain:

- The previous method allows to estimate the weights w_s to apply to each observation n given it belongs to the stratus s.
- In practice this weight has an estimation error ε_s,
- typically when a fraction p_s of the reference population is expected and the observed sample has a size K), the observed fraction is expected to have a variance of $p_s(1-p_s)/K$ that can be propagated in the expression (1).

Writing $\widetilde{w}_n := w_n + \varepsilon_n$ for the weight to apply to observation n, let's **look at what happens when you use the sample in a linear regression**:

$$\min_{\beta} \mathbb{E}_{\widetilde{w}} \|Y - \beta X\|^2 \Rightarrow \hat{\beta} := (X^T \Delta_{\widetilde{w}} X)^{-1} X^T \Delta_{\widetilde{w}} Y.$$

It introduces an **uncertainty on the regression coefficients:** when ε is too high, it may prevent the estimated coefficients to be statistically different from zero...

Take-Aways Of Active Post-Stratification To Correct Biases

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,
  ☞ This is an **active** process
- to adjust weights on your observed sample to get them as close as possible to this external reference.

Open questions

- **the uncertainty due to the size of the sample** has to be taken into account,
- To what extend a sample can be reweigthed? for which application?
  ☞ This should be (re)done each time you have research project.

It is “easy” as long as you restrict youself to groups/categories of observations.

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

$$\mathbb{E}_{(x,y)\sim\mathbb{P}_{\text{te}}}(\ell) = \mathbb{E}_{(x,y)\sim\mathbb{P}_{\text{tr}}}\left(\frac{\mathbb{P}_{\text{te}}}{\mathbb{P}_{\text{tr}}} \cdot \ell\right). \tag{2}$$

You face the usual difficulties:

- 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

$$\begin{aligned} \min \quad & \|\mu(\mathbb{P}_{te}) - \mathbb{E}_{x\sim\mathbb{P}_{tr}}(\beta(x)\Phi(x))\| \\ \text{s.t.} \quad & \beta(x) \geq 0, \\ & \mathbb{E}_{x\sim\mathbb{P}_{tr}}(\beta(x)) = 1 \end{aligned} \tag{3}$$

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**.

☞ It can be reformulated as a **optimal transportation problem**...

Moreover, in practice, setting

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

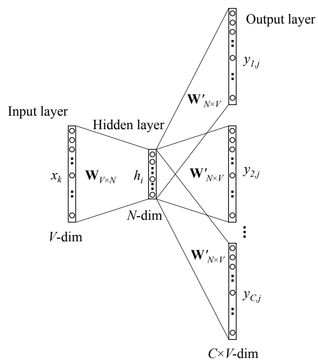
one can write

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

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 the empirical average of β)

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

☞ 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).



Standard word2vec Skip-gram notations of [Rong, 2014]: a document is a sequence $X_1, \ldots, X_k, \ldots, X_T$ of words that belong to a vocabulary $\mathcal{V} := \{x_1, \cdots, x_V\}$ (English vocabulary $\sim$ 70,000 words). The word2vec [Goodfellow et al., 2016] can be written as a neural network with one hidden layer of size N, the embedding size, far lower than V (in practice taken between 200 to 500) and a softmax on the last layer.

In one formula (for the Skip-gram version), if x_i is an input word and x_j an output word, W is the $V \times N$ matrix of embeddings and W' the $N \times V$ matrix of contexts:

$$y^i_{c,j} = \mathrm{softmax}(x_i^T W W') x_j, \qquad \forall c \in \{1, .., C\} \tag{4}$$

is the probability that x_j is in the neighborhood of x_i, here it will mean it is among the C words after x_i.

Nevertheless embeddings are embeddings...

Of course it is possible to be more sophisticated

- You can **use more than one word as input** (it is equivalent to increase the input size)
- You can **"localize" the context and the embeddings** (writing $W(x_i, x_j)$ and $W'(x_i, x_j)$, it is one of the feature of attention heads and it encompasses the use of hidden state variables like topics)
- You can **be more non-linear** (it is not very different than to localize the embeddings).

It is convenient to think that these word2vec embeddings are very close to what is happening locally (around x_i and x_j) in deeper models. Take it as the first term of a Taylor expansion of a BERT...
(does anyone want to prove that?)

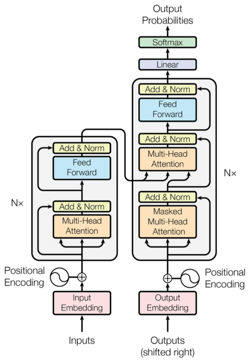


Figure 1: The Transformer - model architecture.

The loss function an one useful definition

The loss function targets to recover a $V \times V$ **stochastic matrix** : row i contains the probabilities to see each word of the vocabulary "around" x_i:

$$\begin{aligned} \ell(X_k, \ldots, X_{k+C}) &= -\log \hat{\mathbb{P}}_{W,W'}(X_{k+1} = x_{j_1}, \cdots, X_{k+C} = x_{j_C} | X_k = x_i) \\ &= -\log \prod_{c=1}^{C} \text{softmax}(x_i^T W W') x_{j_c} \end{aligned} \tag{5}$$

The Reference Model is what embeddings is targeting to recover

A Reference Model RM is a triplet (Id, W_0', C) that can be identified to the parameters of a word2vec neural network. With V the size of the vocabulary, W_0' is a $V \times V$ stochastic matrix which element (i, j) records the probability to see the j-th word of the vocabulary in the C words following the i-th word.

There is a Reference Model corresponding to a word2vec (i.e. $W_0' = WW'$), a Reference Model corresponding to an empirical corpus, and a Reference Model corresponding to the "true" model of the corpus. Reference Models do not need softmax.

Compression? Synonyms?

It is obvious that if you start from a Reference Model W_0' having two similar lines i and i', you can compress it by projecting the two words x_i and $x_{i'}$ on the same one, i.e. by removing row i' of W_0' and using a trivial embedding mapping word i' to word i. This is not the same compression than PCA (more focus on low rank that max variance) [Levy and Goldberg, 2014].

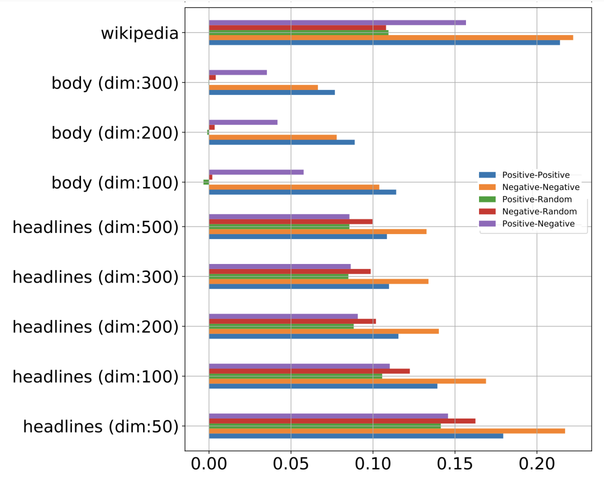
This means that from the viewpoint of embeddings what is important is to recover the probabilistic environment of each word: it is very efficient to map two words having the same environment on the same embedding. Somehow **the loss function $\ell(\cdot)$ seems to push embeddings to associate synonyms to the same embedding**.

☞ Thanks to a Markovian representation of language structure compatible with Language Models, we have been able to build synthetic languages from scratch with desired properties and study how Small Language Models (taken as proxy of LLMs), would misunderstand antonyms as synonyms in some conditions.

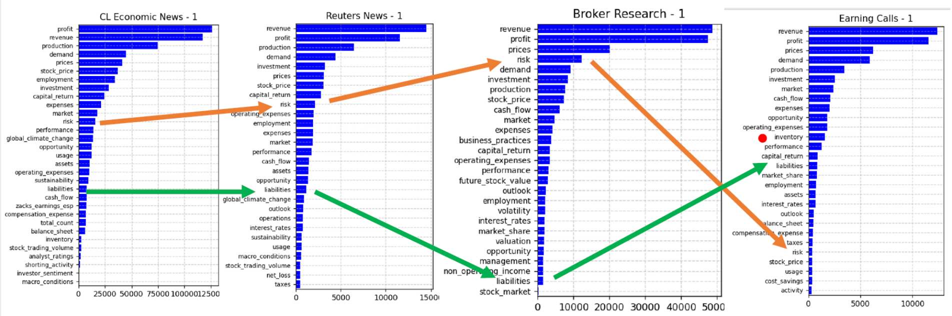
In [Li and Lehalle, 2024, 2021], we compared the proximity between world of a reference polarity in finance [Loughran and McDonald, 2011] using different sizes of embedding.

You want blue and orange bars to be large (similar concepts), red and purple to be low (opposite) and green to be in the middle.

And we found that antonyms can be frequentists synonyms.

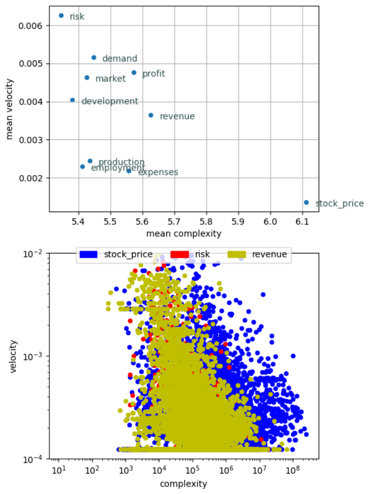


For financial applications, you have specific themes of interest, texts are published on these thematics, addressing specifically some companies.
Quickly you discover that **different corpora of text address different topics**.



Joint work with `CausalityLink`.

Financial Texts As Dynamic Datasets



With CausalityLink, we continued these explorations, based on their capability to identify KPIs in economy and finance (for instance “production”, or “risk”).
It allows to count

- the number of texts issued an a given topic per day: the velocity of the concept,
- the length of and number of other KPI in the text containing an information about this topic, that reflects the complexity of the concept.

☞ A natural ordering of contexts emerges from the data: information on complex concepts cannot be produced too quickly.
It implies that we shouldn’t expect to be updated at an arbitrary frequency on any KPI.

Temporary Conclusion on Texts

The recent wave of Natural Language Processing (NLP)

- started in the nineties with Self Organizing Maps' prototypes [Lagus et al., 1999],
- after a long period of slow progresses in Canada [Mnih and Hinton, 2008]
- made a qualitative progress with Mikolov's celebrated relation on embeddings ("King - Man + Woman = Queen") [Mikolov et al., 2013].

☞ LLM-like architecture have the capability to **use texts as data (in context).**

"But" they are trained on data.

The knowledge these LLMs store is limited and **subject to hallucinations** despite a remarkable capacity to read and speak English (that quickly extended to other languages).
We feed LLMs with gigantic databases of texts without discrimination, probably containing a lot of noise, biases and inaccuracies. More data are now generated programmatically, to inject forged and noiseless knowledge.
Another approach has been to clean or to select more carefully the data [Penedo et al., 2023].

For Financial Application the timing is important.

You would like to monitor narratives ☞ Know and clean your data.

Next Section – The Process Of Taking Decisions

The Process Of Taking Decisions: (linear) Portfolios

A portfolio is simply a linear combination of financial instruments. w are weights (they can change with time), X is a K-dimensional stochastic process of returns, the returns of the portfolio are

$$\mathrm{w}^\top \mathrm{P}\Delta^{1/2}\xi,$$

with the standard PCA notation for $\mathbb{E}\mathrm{X}\mathrm{X}^\top = \mathrm{P}\Delta\mathrm{P}^\top$.

Portfolios are in a lot of places ([Lehalle and Simon, 2021, Bryzgalova et al., 2019] or [Roncalli, 2013]), they are sometimes hidden, like for the estimation of the **Value at Risk** of a portfolio (or an index, that is a portfolio). Indeed the skew of a portfolio often stems from a temporary contraction of the covariances: without looking at the components of the portfolio/index, the estimation of the VaR is probably not correct.

The most common way to use GenAI on financial time series (biblio in 1 slide) is to target a better estimation of the outcome of an maximization "a-la-Markowitz":

$$\begin{array}{ll} \text{maximize} & \mu^\top \mathrm{w} \\ \text{subject to} & \mathrm{w}^\top \Omega \mathrm{w} \leq \mathrm{s}^2, \end{array} \tag{6}$$

where μ are expected returns and Ω a risk model.

☞ With these two slides, we have already a lot of pitfalls...

Reminder On Portfolio Construction

A typical risk model is

(7) $$\Omega = \Sigma + \sigma^2 U, \quad \Sigma = P\Delta P^\top, \quad U = Q\,\mathrm{Id}_{K-d}\,Q^\top,$$

Then for any arbitrary vector z

(8) $$\Omega^{-1} z = \frac{1}{\sigma^2}\left(\sum_{k\le d} \frac{\sigma^2}{\lambda_k} \langle z, P_k\rangle\, P_k + \sum_{k>d} \langle z, Q_k\rangle\, Q_k\right).$$

And the optimal weights w of the Modern Portfolio Optimization program (6), given expected returns $\mu = [P, Q][y_P; y_Q] = [P, Q][y_k]_{1\le k\le K}$ reads

(9) $$\forall k : w_k = \frac{1}{\sigma^2\gamma}\left(\sum_{k\le d} \frac{\sigma^2}{\lambda_k} y_k\, P_k + \sum_{k>d} y_k\, Q_k\right),$$

where γ is the renormalizing constant of (6) to reach the target risk s^2.

Portfolios Can Be Intertemporal

In this framework, time is “flatten” in the sense that the ordering of events is not important.

When you take care of time, i.e. of the trajectories of prices, it allows to measure with accuracy the **Value at Risk** of your portfolio, that is the worst case loss at a given horizon.

Unfortunately, the moment a portfolio suffer from great loss it is when all its components start to move in one or two sole directions: **their correlations increase** (in the previous slide they were constant...).

Only alternative data (typically from the supply chain) can give you clues about what is happening for the fundamentals, and observation of market impact for the liquidity tensions.

As a consequence, **Value at Risk have a temporal shape**, that is encoded in the dynamics of correlation of the underlyings.

Until section seven, it will be assumed that $dy \equiv o$, i.e. all income is derived from capital gains on assets. If one of the n-assets is "risk-free" (by convention, the $n^{\underline{th}}$ asset), then $\sigma_n = o$, the instantaneous rate of return, α_n, will be called r, and (14) is re-written as

(14') $dW = \sum_1^m w_i(\alpha_i - r)Wdt + (rW-C)dt + dy + \sum_1^m W_i\sigma_i dz_i$

In [Merton, 1975] the model is simple and elegant: it takes care of the repartition between risky and risk-free rate assets, and of a level of “consumption” that is the speed at which the investor needs its money back:

Of course, the mathematical formula should not hide we do not really know the long term, even for the risk-free rate [El Karoui et al., 2022] (especially that now we are not sure about our future reference currency...).

Every Position Ends Up In Forming Prices

Some investors and market participants desire to take sophisticated mix of exposure, involving formulas that can be computed only in the future.

For instance: “As a producer of plastic chairs, I would like to be protected against a rise in the oil price (that is my row material), until a point that it will generate so much inflation that I will have less clients.”

Investment banks are structuring and selling these products: **they accept to be accountant of claims in the future**. They can do it only via risk replication: they continuously built the position they will have to deliver.

If I will have to deliver 20 shares to my client when the price will end up above 100: if the price increases to 100, I slowly buy shares; if the price decreases below 100, I slowly sell share.

☞ it is like buying or selling in behalf of my clients, following the most probably outcome: **doing so, I impact the price**. I act as a proxy but the view of my client participate to the double auction game.

Next Section – Moving (Or Forming?) The Price

Moving (Or Forming?) The Price

Once information is transformed in bets, and once bets are combined in portfolios of position, market participants are trading, and **their flows are moving prices**.

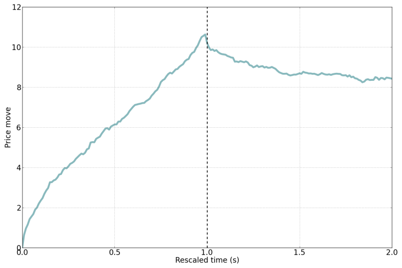
And **prices are moving in conjunction with the information**.

They are two viewpoints, [Bouchaud, 2023]:

- Either price moves are entirely due to information (this is a world without Grossman-Stiglitz),
- Either price moves are 100% coming from market participants actively pushing the price (but in the context of a double-auction game, it does not account from situations where buyers ans sellers agree on a new price only via the medium of information).

☞ the reality is probably in between, but what are the conditioners that would explain the split?

On our database of 300,000 large orders [Bacry et al., 2015]



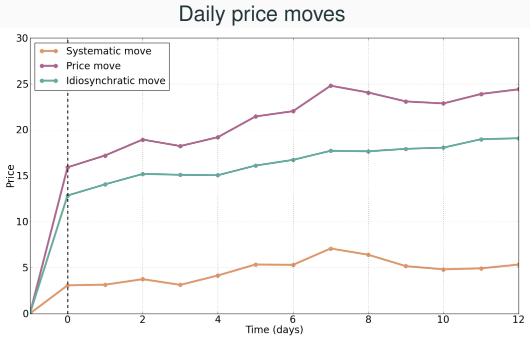
f

Market Impact takes place in different phases

- the **transient impact**, concave in time,
- reaches its maximum, the **temporary impact**, at the end of the metaorder,
- then it **decays**,
- up to a stationary level; the price moved by a **permanent** shift.

Permanent Market Impact

We had enough data to investigate long term impact, exploring the relationships between permanent impact and traded information.



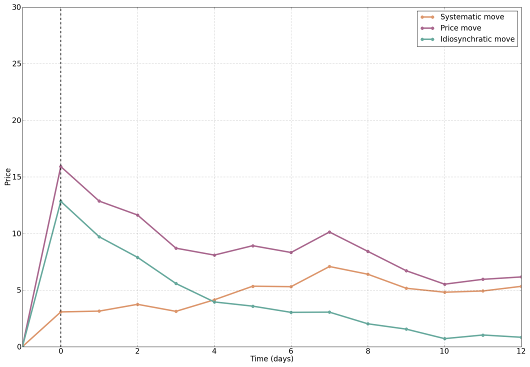
- If you plot the long term price moves (x-axis in days), you see an regular increase;
- But the same stock is traded today, tomorrow, the day after, etc.

👍 This means that permanent market impact is primarily independent of your trading: **if a (good) portfolio manager decides to buy and her systems are down: the permanent market impact will nevertheless be there**.

Permanent Market Impact

We had enough data to investigate long term impact, exploring the relationships between permanent impact and traded information.

Daily price moves



- If you plot the long term price moves (x-axis in days), you see an regular increase;
- But the same stock is traded today, tomorrow, the day after, etc.
- Once you remove the market impact of "future" trades (similarly to [Brokmann et al., 2015]), you obtain a different shape.
- If you look each curve: the yellow one contains the CAPM β (the metaorders are trading market-wide moves), the green curve contains the idiosyncratic moves, this shape can be read as **the daily decay of metaorders impact**.

☞ This means that permanent market impact is primarily independent of your trading: **if a (good) portfolio manager decides to buy and her systems are down: the permanent market impact will nevertheless be there**.

Next Section – Optimal Control Of The Trading Process

Following [Cartea et al., 2015] Define

- the exogenous trading flow μ_t,
- your trading flow ν_t: your control,
- the permanent impact via: $dS = b \cdot (\mu_t - \nu_t)\, dt + \sigma\, dW$,
- the temporary impact via: $\hat{S}_t := S_t - (\psi/2 + \kappa\nu_t)$, where ψ is the bid-ask spread.

Now you have the following dynamics $dQ^\nu = -\nu\, dt$, and $dX^\nu = \hat{S}_t \nu_t\, dt$.

Basic Equation of Cartea-Jaimungal's H

$$0 = (\partial_t + \frac{1}{2}\sigma^2 \partial_S^2 + \mathcal{L}^\mu)H - \phi q^2 + \sup_\nu \left\{ \left(\nu \left[s - (\frac{\psi}{2} + \kappa\nu) \right] \partial_x + b(\mu - \nu)\, \partial_S - \nu\, \partial_q \right) H \right\}.$$

under the terminal condition: $H(T, x, s, \mu, q) = x + q(s - \frac{\psi}{2}) - Aq^2$. Remind A is a large penalization.

Proof: write the Dynamic Programming Principle.

Solving Cartea-Jaimungal's HJB

Replace $H(t,x,s,\mu,q)$ by $x+q(s-\frac{\psi}{2})+h(t,\mu,q)$. You find

$$\partial_t h + \mathcal{L}^\mu h + b\mu q - \phi q^2 + \sup_\nu \left\{ -\kappa\nu^2 - (bq+\partial_q h)\nu \right\} = 0.$$

This is nice since $\nu^* = -(bq+\partial_q h)/(2\kappa)$ is the maximizing control. Injected in the PDE, it reads

$$\partial_t h + \mathcal{L}^\mu h + b\mu q - \phi q^2 + \frac{1}{4\kappa}(bq+\partial_q h)^2 = 0.$$

Assume h is quadratic in q: $h(t,\mu,q) := h_0(t,\mu) + q h_1(t,\mu) + q^2 h_2(t,\mu)$ and collect the terms in h_i:

$$\left\{ \begin{array}{rcl} (\partial_t + \mathcal{L}^\mu) h_0 + \frac{1}{4\kappa} h_1^2 &=& 0 \quad , \quad h_0(T,\mu) = 0 \\ (\partial_t + \mathcal{L}^\mu) h_1 + b\mu + h_1 \frac{1}{2\kappa}(b+2h_2) &=& 0 \quad , \quad h_1(T,\mu) = 0 \\ (\partial_t + \mathcal{L}^\mu) h_2 - \phi + \frac{1}{4\kappa}(b+2h_2)^2 &=& 0 \quad , \quad h_2(T,\mu) = -A \end{array} \right.$$

Solving Cartea-Jaimungal's HJB

1. The third equality shows h_2 is independent of μ: It is a Riccati equation, with solution

$$\boxed{h_2(t) := \xi(t) - \frac{1}{2}b,\ \xi(t) := \sqrt{\kappa\phi}\frac{1+\zeta e^{2\gamma(T-t)}}{1-\zeta e^{2\gamma(T-t)}}}$$

   where $\gamma = \sqrt{\phi/\kappa}$ and $\zeta = (A - b/2 + \sqrt{\phi\kappa})/(A - b/2 - \sqrt{\phi\kappa})$.

2. The second PDE[1] can be solved by Feynman-Fac on μ (with $K(t) := \zeta e^{\gamma(T-t)} - \zeta e^{-\gamma(T-t)}$)

$$\boxed{h_1(t,\mu) := b\int_{u=t}^{T} \frac{K(u)}{K(t)}\mathbb{E}_{t,\mu}(\mu_u)du} = b\,\mathbb{E}_{t,\mu}\int_t^T \exp\left\{\int_t^s u(s)\right\}\mu_s ds.$$

3. The first one $(\partial_t + \mathcal{L}^\mu)h_0 = -\frac{1}{4\kappa}h_1^2$ can be solved by direct integration:

$$\boxed{h_0(t,\mu) := \frac{1}{4\kappa}\int_t^T \left(\mathbb{E}_{t,\mu}h_1^2(t,\mu_u)\right)\,du.}$$

[1] $(\partial_t + \mathcal{L}^\mu)h_1 + b\mu + h_1 \cdot u = 0$ where $u(t) := (b/2 + h_2(t))/\kappa$.

Time To Conclude

Main Take-Aways

Market Participants are in a closed loop of

- **extracting information** from the economic world,
- adjust their decisions by building **portfolio of exposures**,
- the generated flows, in conjunction of **all flows, move prices**.

We have a language to express this a coherent way: it is the language of financial maths, based on trajectory-wise risk analysis of dynamics in uncertain environment, and expressing the decision taking process using the Dynamic Programming Principle.

With big data and machine learning, **information extraction can be formalised with a compatible language** (since ML share the same fundations: stochastic processes, [Benaïm and El Karoui, 2005]).

When you estimate the parameters of a conditional expectation (a prediction) from a dataset to take decisions, you need to understand causality.

Thank you for your attention: Time for more questions

charles-albert.lehalle@polytechnique.edu

Bacry, E., Iuga, A., Lasnier, M., and Lehalle, C.-A. (2015).
Market impacts and the life cycle of investors orders.
Market Microstructure and Liquidity, 1(02):1550009.

BańBura, M., Giannone, D., and Reichlin, L. (2014).
Nowcasting.
In The oxford handbook of economic forecasting.

Benaïm, M. and El Karoui, N. (2005).
Promenade aléatoire: chaînes de Markov et simulations, martingales et stratégies.
Editions Ecole Polytechnique.

Beneish, M. D. (1991).
Stock prices and the dissemination of analysts' recommendation.
Journal of Business, pages 393–416.

Besson, P. and Lehalle, C.-A. (2014).
The deal/book split analysis: A new method to disentangle the contribution to market and limit orders in any price change.
Available at SSRN 2377965.

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

Bouchaud, J.-P. (2023).
Introduction to part ii. price impact: Information revelation or self-fulfilling prophecies?
In Capponi, A. and Lehalle, C.-A., editors, Machine Learning and Data Sciences for Financial Markets: A Guide to Contemporary Practices. Cambridge University Press.

Brokmann, X., Serie, E., Kockelkoren, J., and Bouchaud, J.-P. (2015).
Slow decay of impact in equity markets.
Market Microstructure and Liquidity, 1(02):1550007.

Bryzgalova, S., Pelger, M., and Zhu, J. (2019).
Forest Through the Trees: Building Cross-Sections of Stock Returns.
SSRN Electronic Journal.

Cartea, Á., Jaimungal, S., and Penalva, J. (2015).
***Algorithmic and high-frequency trading*.**
Cambridge University Press.

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.

El Karoui, N., Hillairet, C., and Mrad, M. (2022).
Ramsey rule with forward/backward utility for long-term yield curves modeling.
Decisions in Economics and Finance, 45(1):375–414.

Goodfellow, I., Bengio, Y., and Courville, A. (2016).
***Deep Learning*.**
MIT Press.
http://www.deeplearningbook.org.

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.

Grossman, S. J. and Stiglitz, J. E. (1980).
On the impossibility of informationally efficient markets.
The American economic review, 70(3):393–408.

Kyle, A. S. (1985).
Continuous auctions and insider trading.
Econometrica: Journal of the Econometric Society, pages 1315–1335.

Lagus, K., Honkela, T., Kaski, S., and Kohonen, T. (1999).
Websom for textual data mining.
Artificial Intelligence Review, 13:345–364.

Lehalle, C.-A. and Simon, G. (2021).
Portfolio selection with active strategies: how long only constraints shape convictions.
Journal of Asset Management, 22(6):443–463.

Levy, O. and Goldberg, Y. (2014).
Neural word embedding as implicit matrix factorization.
In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K. Q., editors, Advances in Neural Information Processing Systems, volume 27, pages 2177–2185. Curran Associates, Inc.

Li, M. and Lehalle, C.-A. (2024).
Mathematics of embeddings: Spillover of polarities over financial texts.
Reviews In Modern Quantitative Finance, pages 151–188.

Loughran, T. and McDonald, B. (2011).
When is a liability not a liability? textual analysis, dictionaries, and 10-Ks.
The Journal of finance, 66(1):35–65.

Merton, R. C. (1975).
Optimum consumption and portfolio rules in a continuous-time model.
In Stochastic optimization models in finance, pages 621–661. Elsevier.

Mikolov, T., Yih, W.-t., and Zweig, G. (2013).
Linguistic regularities in continuous space word representations.
In Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies, pages 746–751.

Mnih, A. and Hinton, G. E. (2008).
A scalable hierarchical distributed language model.
Advances in neural information processing systems, 21.

References vi

Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Alobeidli, H., Cappelli, A., Pannier, B., Almazrouei, E., and Launay, J. (2023).
The refinedweb dataset for falcon llm: Outperforming curated corpora with web data only.
Advances in Neural Information Processing Systems, 36:79155–79172.

Roncalli, T. (2013).
***Introduction to risk parity and budgeting*.**
CRC Press.

Rong, X. (2014).
word2vec Parameter Learning Explained.
arXiv e-prints, page arXiv:1411.2738.

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

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