Some contributions to the clustering of financial time series and applications to credit default swaps

Gautier Marti

NNT : 2017SACLX097

Thèse de doctorat de l'Université Paris-Saclay préparée à l'Ecole polytechnique

Ecole doctorale n°580
Sciences et technologies de l'information et de la communication
Spécialité de doctorat : Informatique

par

M. Gautier Marti

Quelques contributions aux méthodes de partitionnement automatique des séries temporelles financières, et applications aux couvertures de défaillance

Thèse présentée et soutenue à l'Ecole polytechnique, le 10 novembre 2017.

Composition du Jury :

Mme	Julie Josse	Professeur Ecole polytechnique	(Présidente)
M.	Damiano Brigo	Professeur Imperial College London	(Rapporteur)
M.	Fabrizio Lillo	Professeur Scuola Normale Superiore	(Rapporteur)
M.	Rama Cont	Professeur Imperial College London	(Examinateur)
M.	Michalis Vazirgiannis	Professeur Ecole polytechnique	(Examinateur)
M.	Fabio Caccioli	Maître de conférences University College London	(Examinateur)
M.	Frank Nielsen	Professeur Ecole polytechnique	(Directeur de thèse)

Contents

[1] reference to the book *Practical Fruits of Econophysics* [203]

Acknowledgements

The content presented in this PhD thesis was elaborated while I was working as a quantitative researcher at Hellebore Capital, an asset management company specialized in credit derivatives arbitrage `www.helleborecapital.com`. I am thus grateful to Hellebore Capital partners for having me hired to carry out this research work.

At Ecole Polytechnique, I worked under the supervision of Frank Nielsen, expert in Computational Information Geometry, who provided during these last three years precious advice.

I also owe much to Philippe Donnat and Philippe Very for ideas and support.

I would like to thank the jury, and especially Prof. Damiano Brigo and Prof. Fabrizio Lillo for their comments and suggestions as rapporteurs of this PhD thesis.

Thank you also to my new employer, AXA IM Chorus, for letting me finish this thesis and defend it in good conditions.

Finally, I am grateful to all current and past colleagues at Hellebore Capital and Technologies, at OTCStreaming, family and friends for having provided me their support in their own way at some point during these industrious years.

Chapter 1

Introduction

1.1 Motivation

Hellebore Capital is investing in developing AI & Machine Learning techniques specifically tailored to financial time series. The goal of Hellebore Capital R&D is to develop models and methodologies that improve upon the body of existing knowledge and, using robust AI & Machine Learning methods could provide an alternative framework in respect of:

- **Risk management**
  - Value-at-Risk
  - SPAN-like methods (stressed scenarios on groups of assets)
- **Investment activity**
  - portfolio design
  - statistical arbitrage
  - market-making
- **Data analysis**
  - cleaning historical time series (missing values imputation, outliers detection)
  - exploring thousands of time series (e.g. `www.datagrapple.com`)

Why are these tasks difficult?

Though AI & Machine Learning have shown impressive success in many applications such as in Computer Vision (self-driving vehicles) and audio signal processing (live translators) during the last 5 years (due in part to the availability of Big Data to train the models), their application to financial time series (such as those displayed in Figure 1.1) is much more involved for several reasons:

1. non-stationarity
   since the conditions are always changing it is hard to have lots of data and it would be misleading to use data from the distant past for fitting the models;

2. near efficiency
   the financial time series of asset prices behave nearly like random walks, no trivial patterns are to be found;

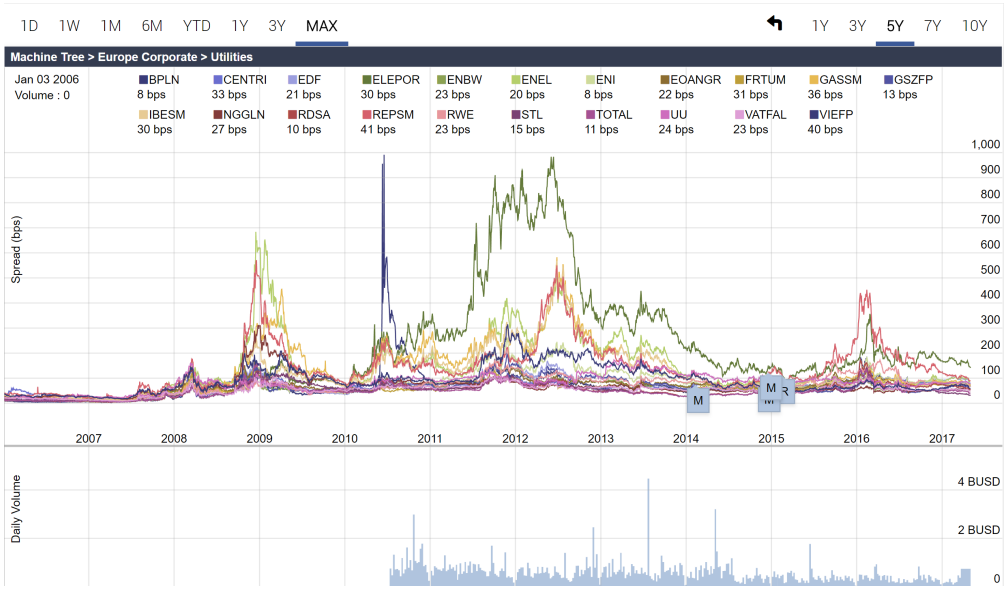


Figure 1.1: Time series of spreads (bps) for European corporates

3. very low signal-to-noise ratio
   these time series are very noisy (non-relevant information and measure artifacts hide information in random fluctuations) and overfitting is a common pitfall with this kind of data; and

4. unfavorable statistical setting
   having many time series to study but only few observations for each of them (due to reason 1 above), it makes the estimation of many quantities (such as a correlation matrix) even more challenging (adding to reason 3 above).

For these reasons, fitting robust models is hard and out-of-sample results are often poor.

Finding groups of assets who share a similar behaviour helps to reduce dimensionality and thus alleviates the problems mentioned above. Clusters, i.e. groups of similarly behaving assets, can be important building blocks for bigger systems. The present work aims at grounding their construction and use in order to obtain more robust models.

1.2 Main contributions and outline of the thesis

In Chapter 2, we start this thesis by providing an extensive review of the field of research. The relevant literature is scattered among different research fields. Econophysics is the major one, but one can find useful contributions from the following literature: econometrics, finance and accounting, risk and quantitative finance, multivariate analysis, bayesian statistics, data mining, intelligent and fuzzy systems, machine learning and artificial intelligence, etc. We also take advantage of this review to contextualize our main contributions that are listed in the Bibliography below.

In Chapter 3, we present the datasets built and maintained by Hellebore Capital which both motivated and showcased the methodologies we developed. We also take advantage of this chapter to convey key insights into the credit default swap markets.

The core part of the thesis, i.e. Chapters 4 & 5, is dedicated to highlight our main published contributions. The Chapter 6 provides practical methods for a good clustering analysis of the financial time series, e.g. model selection, sanity checks, visualization, etc.

We conclude by providing some avenues for further research and open questions which are ongoing research problems.

Bibliography

[1] Philippe Donnat, Gautier Marti, and Philippe Very. Toward a generic representation of random variables for machine learning. *Pattern Recognition Letters*, 70:24–31, 2016.

[2] Gautier Marti, Sébastien Andler, Frank Nielsen, and Philippe Donnat. Clustering financial time series: How long is enough? In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pages 2583–2589, 2016.

[3] Gautier Marti, Sébastien Andler, Frank Nielsen, and Philippe Donnat. Exploring and measuring non-linear correlations: Copulas, lightspeed transportation and clustering. In *Proceedings of the NIPS 2016 Time Series Workshop, co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016*, pages 59–69, 2016.

[4] Gautier Marti, Frank Nielsen, Philippe Donnat, and Sébastien Andler. On clustering financial time series: a need for distances between dependent random variables. In *Computational Information Geometry*, pages 149–174. Springer, 2017.

[5] Gautier Marti, Frank Nielsen, and Philippe Donnat. Optimal copula transport for clustering multivariate time series. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*, pages 2379–2383, 2016.

[6] Gautier Marti, Sébastien Andler, Frank Nielsen, and Philippe Donnat. Optimal transport vs. fisher-rao distance between copulas for clustering multivariate time series. In *IEEE Statistical Signal Processing Workshop, SSP 2016, Palma de Mallorca, Spain, June 26-29, 2016*, pages 1–5, 2016.

[7] Gautier Marti, Philippe Very, Philippe Donnat, and Frank Nielsen. A proposal of a methodological framework with experimental guidelines to investigate clustering stability on financial time series. In *14th IEEE International Conference on Machine Learning and Applications, ICMLA 2015, Miami, FL, USA, December 9-11, 2015*, pages 32–37, 2015.

[8] Gautier Marti, Frank Nielsen, Philippe Very, and Philippe Donnat. Clustering random walk time series. In *Geometric Science of Information - Second International Conference, GSI 2015, Palaiseau, France, October 28-30, 2015, Proceedings*, pages 675–684, 2015.

Part I

Some background on the thesis challenges: scattered state of the art and over-the-counter credit default swaps

Chapter 2

A review of two decades of correlations, hierarchies, networks and clustering in financial markets

This chapter corresponds to a review that is still evolving and that I shall try to keep it updated with new research and findings. The working document can be found on arXiv: `https://arxiv.org/abs/1703.00485` [138]. We highlight in blue our contributions.

2.1 The standard and widely adopted methodology

The methodology which is widely adopted in the literature stems from Mantegna's seminal paper [133] (cited more than 1200 times as of 2017) and chapter 13 of the book [134] (cited more than 3600 times as of 2017) published in 1999. We describe it below:

- Let N be the number of assets.
- Let $P_i(t)$ be the price at time t of asset i, $1 \leq i \leq N$.
- Let $r_i(t)$ be the log-return at time t of asset i:

$$r_i(t) = \log P_i(t) - \log P_i(t-1).$$

- For each pair i, j of assets, compute their correlation:

$$\rho_{ij} = \frac{\langle r_i r_j \rangle - \langle r_i \rangle \langle r_j \rangle}{\sqrt{\left(\langle r_i^2 \rangle - \langle r_i \rangle^2\right)\left(\langle r_j^2 \rangle - \langle r_j \rangle^2\right)}}.$$

- Convert the correlation coefficients ρ_{ij} into distances:

$$d_{ij} = \sqrt{2(1-\rho_{ij})}.$$

- From all the distances d_{ij}, compute a minimum spanning tree (MST) using, for example, Algorithm 1:

  Several other algorithms are available to build the MST [98].

The methodology described above builds a tree, i.e. a connected graph with $N-1$ edges and no loop. This tree is unique as soon as all distances d_{ij} are different. The resulting MST also provides a unique indexed hierarchy [134] which corresponds to the one given by the dendrogram obtained using the Single Linkage Clustering Algorithm.

Algorithm 1 Kruskal's algorithm

procedure BUILDMST($\{d_{ij}\}_{1 \leq i,j \leq N}$)
  ▷ Start with a fully disconnected graph $G = (V, E)$
  $E \leftarrow \emptyset$
  $V \leftarrow \{i\}_{1 \leq i \leq N}$
  ▷ Try to add edges by increasing distances
  for $(i, j) \in V^2$ ordered by increasing d_{ij} **do**
    ▷ Verify that i and j are not already connected by a path
    if not *connected*(i, j) **then**
      ▷ Add the edge (i, j) to connect i and j
      $E \leftarrow E \cup \{(i, j)\}$
  ▷ G is the resulting MST
  return $G = (V, E)$

2.2 Methodological concerns and extensions

2.2.1 Concerns about the standard methodology

We list below the concerns that have been raised about the standard methodology during the last 20 years:

- The clusters obtained from the MST (or equivalently, the Single Linkage Clustering Algorithm (SLCA)) are known to be **unstable** (small perturbations of the input data may cause big differences in the resulting clusters) [137].
- The clustering instability may be partly due to the algorithm (MST/Single Linkage are known for the **chaining phenomenon** [49]).
- The clustering instability may be partly due to the correlation coefficient (**Pearson linear correlation**) defining the distance which is known for being **brittle to outliers**, and, more generally, not well suited to distributions other than the Gaussian ones [67].
- **Theoretical results** providing the statistical reliability of hierarchical trees and correlation-based networks are still **not available** [215].
- One might expect that the higher the correlation associated to a link in a correlation-based network is, the higher the **reliability** of this link is. In [220], authors show that this is **not always observed empirically**.
- Changes affecting specific links (and clusters) during prominent crises are of **difficult interpretation** due to the **high level of statistical uncertainty** associated with the correlation estimation [195].
- The standard method is somewhat **arbitrary**: A change in the method (e.g. using a different clustering algorithm or a different correlation coefficient) may yield a huge change in the clustering results [123], [137]. As a consequence, it implies huge variability in portfolio formation and perceived risk [123].

Notice that Benjamin F. King in his 1966 paper [112] (the first paper, to the best of our knowledge, about clustering stocks based on their historical returns; apparently unknown to Mantegna and his colleagues who reinvented a similar method) adds a final footnote which serves both as an advice and a warning for future work and applications:

> One final comment on the method of analysis: this study has employed techniques that rely on finite variances and stationary processes when there is considerable doubt about the existence of these conditions. It is believed that a convincing argument has been made for acceptance of the hypothesis that a small number of factors, market and industry, are sufficient to explain the essential comovement of a large group of stock prices; it is possible, however, that more satisfactory results

could be obtained by methods that are distribution free. Here we are thinking of a factor-analytic analogue to median regression and non-parametric analysis of variance, where the measure of distance is something other than expected squared deviation. In future research we would probably seriously consider investing some time in the exploration of distribution free methods.

It is only but recently that researchers have started to focus on these shortcomings as we will observe through the research contributions detailed in the next section.

2.2.2 Contributions for improving the methodology

To alleviate some of the shortcomings mentioned in the previous section, researchers have mainly proposed alternative algorithms and enhanced distances. Some refinements of the methodology as a whole, alongside efforts to tackle the concerns about statistical soundness, have been proposed.

On algorithms

Several alternative algorithms have been proposed to replace the minimum spanning tree and its corresponding clusters:

- **Average Linkage Minimum Spanning Tree (ALMST)** [220]; Authors introduce a spanning tree associated to the Average Linkage Clustering Algorithm (ALCA); It is designed to remedy the unwanted chaining phenomenon of MST/SLCA.
- **Planar Maximally Filtered Graph (PMFG)** [218] which strictly contains the Minimum Spanning Tree (MST) but encodes a larger amount of information in its internal structure.
- **Directed Bubble Hierarchal Tree (DBHT)** [196] which is designed to extract, without parameters, the deterministic clusters from the PMFG.
- Clustering using **Potts super-paramagnetic transitions** [118]; When anti-correlations occur, the model creates repulsion between the stocks which modify their clustering structure.
- Clustering using **maximum likelihood** [90, 89]; Authors define the likelihood of a clustering based on a simple 1-factor model, then devise parameter-free methods to find a clustering with high likelihood.
- Clustering using **Random Matrix Theory (RMT)** [171]; Eigenvalues help to determine the number of clusters, and eigenvectors their composition.
- Clustering using the p-**median problem** [115]; With this construction, every cluster is a star, i.e. a tree with one central node.

On distances

At the heart of clustering algorithms is the fundamental notion of distance that can be defined upon a proper representation of data. It is thus an obvious direction to explore. We list below what has been proposed in the literature so far:

- Distances that try to quantify how one financial instrument provides information about another instrument:
  - Distance using **Granger causality** [18],
  - Distance using **partial correlation** [107],
  - Study of asynchronous, **lead-lag relationships by using mutual information** instead of Pearson's correlation coefficient [81, 180],

  - The correlation matrix is normalized using the affinity transformation: the correlation between each pair of stocks is normalized according to the correlations of each of the two stocks with all other stocks [108].
- Distances that aim at including non-linear relationships in the analysis:
  - Distances using mutual information, mutual information rate, and other **information-theoretic** distances [82, 180, 8],
  - The Brownian distance [229],
  - **Copula-based** [136], [71, 26] and **tail dependence** [74] distances.
- Distances that aim at taking into account multivariate dependence:
  - Each stock is represented by a bivariate time series: its returns and traded volumes [29]; a distance is then applied to an *ad hoc* transform of the two time series into a symbolic sequence,
  - Each stock is represented by a multivariate time series, for example the daily (high, low, open, close) [121]; Authors use the **Escoufier's RV coefficient** (a multivariate extension of the Pearson's correlation coefficient).
- A distance taking into account both the **correlation** between returns **and** their **distributions** [67].

On other methodological aspects

Besides research contributions on algorithms and distances, other methodological aspects have been pushed further.

- Reliability and statistical uncertainty of the methods:
  - A **bootstrap** approach is used to estimate the statistical reliability of both hierarchical trees [216], [139] and correlation-based networks [220],
  - **Consistency** proof of clustering algorithms for recovering clusters defined by nested block correlation matrices; Study of empirical **convergence rates** [139],
  - **Kullback-Leibler divergence** is used to estimate the amount of filtered information between the sample correlation matrix and the filtered one [217],
  - **Cophenetic correlation** is used between the original correlation distances and the hierarchical cluster representation [166],
  - Several measures between successive (in time) clusters, dendrograms, networks are used to estimate **stability** of the methods, e.g. cophenetic correlation between dendrograms in [165], adjusted Rand index (ARI) between clusters in [137], mutual information (MI) of link co-occurrence between networks in [195].
- Preprocessing of the time series:
  - **Subtract the market mode** before performing a cluster or network analysis on the returns [22],
  - Encode both **rank statistics** and a **distribution histogram** of the returns into a **representative vector** [67],
  - Fit an ARMA(p,q)-FIEGARCH(1,d,1)-cDCC process (**econometric preprocessing**) to obtain dynamic correlations instead of the common approach of rolling window Pearson correlations [188],
  - Use a clustering of successive correlation matrices to **infer a market state** [166].
- Use of **other types of networks**: threshold networks [160], influence networks [86], partial-correlation networks [107, 106], Granger causality networks [18, 225], cointegration-based networks [211], bipartite networks [221], etc.

- Understanding of the drivers of synchronous correlations using the properties of the collective stock dynamics at **shorter time scales** [58] by using **directed networks** of lagged correlations [58, 57].

2.3 Dynamics of correlations, hierarchies, networks and clustering

Many of the empirical studies are based on the whole period available from the data. Some researchers have started to investigate the dynamics of the empirical correlations, and also the hierarchies, networks and clusters extracted from them (cf. [163] as one of the earliest work). This dynamic setting which has the potential to track changes of the market structure is more interesting for practitioners (e.g. risk managers, traders, regulatory agencies). This research is still in its infancy and we think its results are still hardly exploitable in practice. For instance, an interesting but difficult question is the following: Are **changes** in the correlation structure due to statistical **noise** and data artifacts **or** do they provide a **real signal**?

No predominant methodology has emerged for now but the naive one which consists in:

- Computing Pearson correlations on a **rolling window of arbitrary length**,
- then computing a network or a clustering based on the rolling empirical correlation matrix.

Besides the shortcomings of Pearson correlation detailed above, this approach is brittle due to its strong **dependence *a priori* on:**

- the **sampling frequency** (e.g., intraday, daily, weekly),
  - Concerning the sampling frequency, authors in [20] notice that at intraday frequency level some time is needed before the cluster organization emerges completely. According to the paper, "the changes observed in the structure of the MST and of the hierarchical tree suggest that the intrasector correlation decreases faster than intersector correlation between pairs of stocks" when sampling frequency increases. In [22], [137], authors observe that the clusters obtained using daily returns are similar to the ones obtained with weekly timescales, and even to some extent to the ones using monthly returns. Most of the empirical studies focus on daily returns and only a few explore intraday data: [20, 102, 22, 230, 122, 58]. Working with higher frequencies (e.g. at the transaction or quote level) brings further difficulties such as coping with asynchronous data and the Epps effect [79].
- the **length** T of the rolling window,
  - What is the right length for the rolling window? No clear-cut answer has yet been proposed and, in most studies, its length is set somewhat arbitrarily. In [163], authors posit that "the choice of window width is a trade-off between too noisy and too smoothed data for small and large window widths, respectively" and that they "have explored a large scale of different values for both parameters, and the given values were found optimal". What are the proper criteria for setting the window length? The choice can be driven by the goal (e.g. time investment horizon), by regulatory rules (e.g. computing Value-at-Risk using 1-year historical data), by the stability of clusters [137], by a statistical convergence rate [139], by economic regimes or by a trade-off of the preceding criteria.
- the **number** N **of assets** studied.

- The number of considered assets has also a significant impact on the results: the ratio T/N drives the precision of correlation estimation and ultimately the clustering [23], [139, 48].

This dependence makes it difficult to fully understand and analyze results. Once these 'parameters', i.e. the sampling frequency, T, and N, are chosen, one can study

- the **dynamics of correlations**:
  - In [108], authors are using a sliding window of $T = 22$ days to measure and monitor the eigenvalue entropy of the stock correlation matrices (estimated using daily returns, for $N = 25$ (Tel-Aviv stock market), and $N = 455$ (from S&P500)). They also propose a 3D visualization to monitor the configuration of stocks using a 3D PCA.
  - [148] notices three regime shifts during the period 1989-2011 by monitoring eigenvalues and eigenvectors of the empirical correlation matrices (estimated using quarterly recorded prices from the US housing market; $T = 60$, $N = 51$, the number of US states).
- the **dynamics of the MST** and other hierarchical trees:
  Using summary statistics:
  - The MST which evolves over time is monitored using summary statistics (also called topological features) [161] such as the normalized tree length [163], the mean occupation layer [163], the tree half-life [163], a survival ratio of the edges [162, 102, 188], node degree, strength [188], eigenvector, betweenness, closeness centrality [188], the agglomerative coefficient [145].
  - Using these statistics, [163] notices that:
    * the MST strongly shrinks during a stock market crisis,
    * the optimal Markowitz portfolio lies practically at all times on the outskirts of the tree,
    * the normalized tree length and the investment diversification potential are very strongly correlated.
  - And [188] notices that in the Asia-Pacific stock market:
    * the DST (dynamic alternative of the MST, built from dynamic correlations) shrinks over time,
    * Hong Kong is found to be the key financial market,
    * the DST has a significantly increased stability in the last few years,
    * the removal of the key player has two effects: there is no clear key market any longer and the stability of the DST significantly decreases.
  - In [103], authors observe that for the Japanese and Korean stock markets, there is a decrease of grouping by industry categories.

  Using distances or similarity measures between successive dendrograms:
  - Cophenetic correlation coefficient. In [145], authors propose a cophenetic analysis of public debt dendrograms in the European Union ($N = 29$ countries) computed using Pearson correlation of quarterly debt-to-GDP ratios between 2000 Q1 and 2014 Q1 ($T = 57$) with a sliding window of size $w = 15$.
- the **dynamics of clusters**:
  - The paper [115] finds that the cluster structures are more stable during crises (using the p-median problem, an alternative clustering methodology).
  - Authors in [102] notice that there is an "ecology of clusters": They "can survive for finite periods of time during which time they may evolve in some identifiable way before eventually dissipating or dying".

– In [166], the authors track the merging, splitting, birth, death, contraction, and growth of the clusters in time.

2.4 Financial applications

Though many of the academic studies focus on the MST or the clusters *per se*, some papers try to extend their use beyond the filtering of empirical correlation matrices. It has been proposed to leverage them for making financial policies, optimizing portfolios, computing alternative Value-at-Risk measures, etc.

2.4.1 Portfolio design and trading strategies

- [163] finds that the Markowitz portfolio layer in the MST is higher than the mean layer at all times.
- As the stocks of the minimum risk portfolio are found on the outskirts of the tree [174, 163], authors expect larger trees to have greater diversification potential.
- In [207, 165], authors compare the Markowitz portfolios from the filtered empirical correlation matrices using the clustering approach, the RMT approach and the shrinkage approach.
- [177, 169] propose to invest in different part of the MST depending on the estimated market conditions.
- It appears that a large number of stocks are unnecessary for building an index of market change [112].
- The paper [69] describes methods for index tracking and enhanced index tracking based on clusters of financial time series.
- [74] introduces a procedure to design portfolios which are diversified in their tail behavior by selecting only a single asset in each cluster.
- [7] investigates several network and hierarchy based active portfolio optimizations, and find their out-of-sample performance competitive with respect to conventional ones.
- In [166], they suggest that tracking the merging, splitting, birth, and death of the clusters in time could be the basis for pairs-like reversal trading strategies but with pairs corresponding to clusters.
- Earnings per share forecasts prepared on the basis of statistically grouped data (clusters) outperform forecasts made on data grouped on traditional industrial criteria as well as forecasts prepared by mechanical extrapolation techniques [78].
- [186] suggests that one may design a new set of Ricci network curvature based-strategies in statistical arbitrage (e.g. for mean-reverting portfolios).

2.4.2 Risk management

How much money a given portfolio can lose? in normal market conditions? in stressed market conditions? in the presence of systemic risk?

To answer these questions, the use of clusters and networks can help. As presented previously, the clustering hierarchy can be used to filter a correlation [207, 213] or a tail dependence [74] matrix, which helps to measure the risk in normal and stressed market conditions respectively. The systemic risk as defined by the Bank for International Settlements is the risk that a failure of a participant to meet its contractual obligations may in turn cause other participants to default, with the chain reaction leading to broader financial difficulties. Networks seem thus a particularly relevant tool to study this kind of risk.

- Study of **systemic risk**:

- In [94], authors assert that the diminution of regulation has removed barriers between sectors and regions allowing bank to diversify their risk, but it also increased the economic risk through increased interdependencies.
- The paper [120] is focused on energy derivative markets, and their market integration which can be seen as a necessary condition for the propagation of price shocks. The MST is used to "identify the most probable and the shortest path for the transmission of price shocks".
- Authors in [148] focus on the US housing market. According to the paper, "dramatic increases in the systemic risk are usually accompanied by regime shifts, which provide a means of early detection of housing bubbles." They find a sharp increase in housing market correlations over the past decade, indicating that systemic market risk has also greatly increased; They observe that prices diffuse in complex ways that do not require geographical clusters unlike worldwide stock markets which exhibit clear geographical clustering [195].
- The paper [229] is focused on the shipping market. Authors explore the connections between the shipping market and the financial market: The shipping market can provide efficient warning before market downturn. Alike many economic systems which have been exhibiting an increase in the correlation between different market sectors, a factor that exacerbates the level of systemic risk, the three major world shipping markets, (i) the new ship market, (ii) the second-hand ship market, and (iii) the freight market, have experienced such an increase. Authors show it using the MST, Granger causality analysis, and Brownian distance on the prices of the real shipping market, and the stock prices of publicly-listed shipping companies.
- [18] investigates the monthly returns of hedge funds, banks, broker/dealers, and insurance companies. They find that all four sectors have become highly interrelated over the past decade, likely increasing the level of systemic risk.
- [186] shows that Ricci curvature may serve as an indicator of fragility in the context of financial networks.
- [166] detects distinct correlation regimes between 1998 and 2013. These correlation regimes have been significantly different since the financial crisis of 2008 than they had been previously. Cluster tracking shows that asset classes are now less separated. Correlation networks help the authors to identify "risk-on" and "risk-off" assets.
- In [154], authors study the clusters' composition evolution, and their persistence. They observe that the clustering structure is quite stable in the early 2000s becoming gradually less persistent before the unfolding of the 2007-2008 crisis. The correlation structure eventually recovers persistence in the aftermath of the crisis, settling up a new phase which is distinct from the pre-crisis structure one, where the market structure is less related to industrial sector activity.

  [99] finds that financial institutions which have, in the correlation networks, greater node strength, larger node betweenness centrality, larger node closeness centrality and larger node clustering coefficient tend to be associated with larger systemic risk contributions.

- Risk management **methods**:
  - In [72], authors design clusters that tend to be comonotone in their extreme low values: To avoid contagion in the portfolio during risky scenarios, an investor should diversify over these clusters.
  - As far as diversification is concerned, portfolio managers should probably focus on the most stable parts of the graph [120].

– In [152], authors postulate the existence of a hierarchical structure of risks which can be deemed responsible for both stock multivariate dependency structure and univariate multifractal behaviour, and then propose a model that reproduces the empirical observations (entanglement of univariate multi-scaling and multivariate cross-correlation properties of financial time series).

We found that the risk literature using correlation networks and clusters consists essentially in descriptive studies. For now, there are only too few propositions in the academic literature to build effective network-based or cluster-based risk systems.

2.4.3 Financial policy making

Clusters and networks can help designing financial policies. Several papers propose to leverage them to detect risky market environments, develop indicators that can predict forthcoming crisis or economic recovery [230], or find key markets and assets that drive a whole region, and on which stimulus can be applied effectively.

- Authors of [94] claim that "separation prevents failure propagation and connections increase risks of global crises" whereas the prevailing view in favor of deregulation is that banks, by investing in diverse sectors, would have greater stability. To support their argument, using financial networks, they study the aftermath of the Glass-Steagall Act (1933) repeal by Clinton administration in 1999. They find that erosion of the Glass-Steagall Act, and cross sector investments eliminated "firewalls" that could have prevented the housing sector decline from triggering a wider financial and economic crisis:

  > Our analysis implies that the investment across economic sectors itself creates increased cross-linking of otherwise much more weakly coupled parts of the economy, causing dependencies that increase, rather than decrease, risk.

- According to [18], bank and insurance capital requirements and risk management practices based on VaR, which are intended to ensure the soundness of individual financial institutions, may amplify aggregate fluctuations if they are widely adopted:

  > For example, if the riskiness of assets held by one bank increases due to heightened market volatility, to meet its VaR requirements the bank will have to sell some of these risky assets. This liquidation may restore the bank's financial soundness, but if all banks engage in such liquidations at the same time, a devastating positive feedback loop may be generated unintentionally. These endogenous feedback effects can have significant implications for the returns of financial institutions, including autocorrelation, increased correlation, changes in volatility, Granger causality, and, ultimately, increased systemic risk, as our empirical results seem to imply.

- In [120], authors find that the move towards integration started some time ago and there is probably no way to stop or refrain it. However, regulation authorities may act in order to prevent prices shocks from occurring, especially in places where their impact may be important.

2.5 Practical fruits of clusters, networks, and hierarchies[1]

2.5.1 Stylized facts

Stylized facts can be described as follows [54]:

[1] reference to the book *Practical Fruits of Econophysics* [203]

> A set of [statistical] properties, common across many instruments, markets and time periods, [which] has been observed by independent studies.

From the papers we reviewed, we can list the following stylized facts:

- Elements belonging to some economic sectors are strongly connected within themselves, whereas others are much less connected.
- The Energy and Financial sectors are examples of strong connections whereas elements belonging to the Conglomerates, Consumer cyclical, Transportation, and Capital Goods sectors are weakly connected.
- General Electric is at the center of US stocks networks (for several centrality criteria) [133, 20, 163, 29].
- The Energy, Technology, and Basic Materials sectors are sectors of elements significantly connected among them but weakly interacting with stocks belonging to different economic sectors.
- The Financial sector is strongly connected within, but also to others.
- The assets of the classic Markowitz portfolio are always located on the outer leaves of the tree [163].
- The maximum eigenvalue of the correlation matrix, which carries most of the correlations, is very large during market crashes [70] (increased value of the mean correlation).
- The MST shrinks during market crashes [163] and contains a low number of clusters [145].
- The MST provides a taxonomy which is well compatible with the sector classification provided by an outside institution [134, 163].
- Scale free (i.e. the degree of vertices is power law distributed $f(n) \sim n^{-\alpha}$) structure of the MST [222, 111, 163, 21], but the scaling exponent depends on market period and window width [162].
- The MST obtained with the one-factor model is very different from the one obtained using real data [21]. This invalidates the Capital Asset Pricing Model which is based on the one-factor model $r_i(t) = \alpha_i + \beta_i r_M(t) + \epsilon_i(t)$.
- Stocks compose a hierarchical system progressively structuring as the sampling time horizon increases [219, 22].
- The correlation among market indices presents both a fast and a slow dynamics. The slow dynamics is a gradual growth associated with the development and consolidation of globalization. The fast dynamics is associated with events that originate in a specific part of the world and rapidly (in less than 3 months) affect the global system [195, 148].
- Removing the dynamics of the center of mass decreases the level of correlations, but also makes the cluster structure more evident [22].
- Scale invariance of correlation structure (by subtraction of the market mode) might have important implications for risk management, because it suggests that correlations on short time scales might be used as a proxy for correlations on longer time-horizons [22].
- The MST is star-like in low-volatility segments, and chain-like in high-volatility segments [230].
- Volatility shocks always start at the fringe and propagate inwards [230].
- The "post-subprime" regime correlation matrix shows markedly higher absolute correlations than the others [166].
- In [166], authors find far less asset class separation in the post-subprime period.
- One can distinguish three types of topological configurations for the companies: (i) important nodes, (ii) links and (iii) dangling ends [222].
- A node keeps the majority of its neighbours. The non-randomness of the stock market topology is thus a robust property [222].

- The largest eigenvector of the correlation matrix is strongly non-Gaussian, tending to uniform - suggesting that all companies participate. Authors find indeed that all components participate approximately equally to the largest eigenvector. This implies that every company is connected with every other company. In the stock market problem, this eigenvector conveys the fact that the whole market "moves" together and indicates the presence of correlations that pervade the entire system [171].
- The measure of the average length of shortest path in the PMFG shows a small world effect present in the networks at any time horizon [219].
- Among the 100 largest market capitalization stocks in the NYSE, the auto and lagged intraday correlations play a much more prominent role in 2011-2013 than in 2001-2003 [58].
- Authors in [58] find striking periodicities in the validated lagged correlations, characterized by surges in network connectivity at the end of the trading day.
- At short time scales, measured synchronous correlations among stock returns tend to be lower in magnitude [79], but lagged correlations among assets may become non-negligible [209, 57].
- Banks may be of more concern than hedge funds from the perspective of connectedness [18].
- A lack of distinct sector identity in emerging markets [164]; Few largest eigenvalues deviate from the bulk of the spectrum predicted by RMT (far fewer than for the NYSE) [164].
- Emergence of an internal structure comprising multiple groups of strongly coupled components is a signature of market development [164].

2.5.2 Moot points and controversies

Though most of the conclusions of empirical studies do agree, we find some claims that seem to be contradictory:

- Volatility shocks always start at the fringe and propagate inwards [230], but in [194], authors assert that the credit crisis spreads among affected stocks from more centralized to more outer ones, as spread the news about the extent of damage to the global economy.
- One might expect that the higher the correlation associated to a link in a correlation-based network is, the higher the reliability of the link is. The paper [220] shows that it is not always observed empirically. However, the Cramér-Rao lower bound (CRLB) for correlation [144] points out that the higher the correlation, the easier its estimation, i.e. less statistical uncertainty for high correlations.
- For filtering the correlation matrix, SLCA is more stable than ALCA according to [215], but ALCA is more stable and appropriate than SLCA according to [207, 166].
- During a crisis period, is there an increase or decrease of clusters stability? Most papers find a decrease (e.g. [120, 154]), but at least one [115] (using an alternative clustering methodology, the p-median problem) advocates for an increase.

Chapter 3

Introduction to credit default swaps

3.1 Credit default swaps

This chapter does not intend to give a thorough presentation of credit default swaps and it is not the main intent of this thesis. However, we feel that it is important to give the reader some flavour of the credit default swap market and its data mechanism since it drove much of the modeling decisions and tools presented in the remaining part of the thesis. We feel confident though that the techniques we developed can be applied to many other time series sharing similar characteristics such as meteorological time series (e.g. temperature, rainfall) and economic variables (e.g. Gross Domestic Product, Consumer Price Index, unemployment).

3.1.1 A short introduction to credit default swaps

The credit default swap market is an important one. Its outstanding notional amounts to trillions of dollars, often exceeding the gross domestic product of nations.

A credit default swap (CDS) is a financial swap agreement that the seller of protection will compensate the buyer of protection in the case of a credit event such as default of the underlying for example (cf. Figure 3.1 from Wikipedia article 'Credit default swap'). This bilateral contract which transfers the credit risk of a specific company or sovereign from one party to another for a specified period of time can be considered as the simplest and most important credit derivative [159]: CDS is like an insurance on the default of the underlying (usually, but not exclusively, a corporate bond). Credit default swaps are thus a very important financial product that banks can use to hedge counterparty credit risk and credit valuation adjustment (CVA) [35].

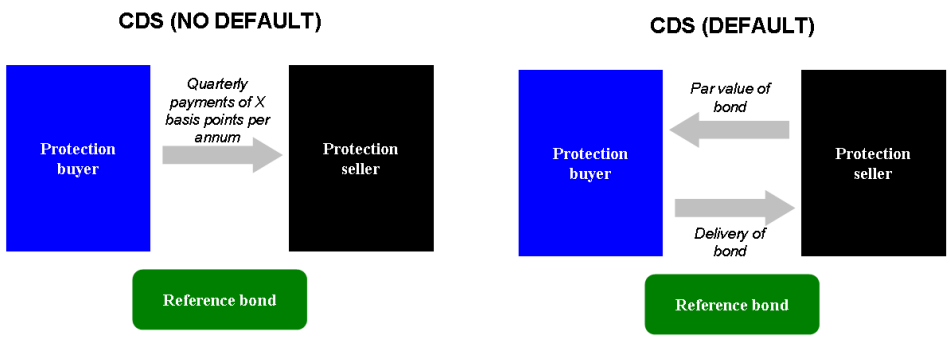


Figure 3.1: Left: No default occurs during the lifetime of the CDS, the protection seller gets his quarterly coupons until maturity of the contract; Right: A default occurs before the maturity of the CDS, the protection seller stopped being paid and must pay the protection buyer the notional of the CDS minus recovery (the figure describes an hedging use case: the CDS notional is chosen to be the par value of a bond)

The precise definitions of credit events and credit default swap contracts are given by the International Swaps and Derivatives Association (ISDA) `http://www2.isda.org/`. One of the ISDA's mission is to standardize and document the market structure of these over-the-counter (OTC) derivatives products.

3.1.2 CDS pricing

The main idea for pricing a credit default swap (and other derivatives) is that it should be a fair contract at inception: neither of the two counterparties should have an advantage. It means that the present value, i.e. the sum of the expected discounted cash flows (depicted in Figure 3.2), of the protection-buyer leg has to be equal to the present value of the protection-seller leg.

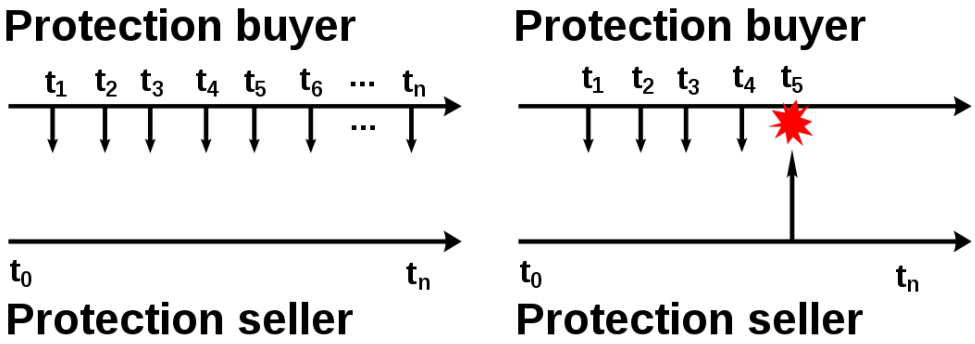


Figure 3.2: Left: In case of 'no default' the protection buyer pays to the protection seller quarterly coupons. Before the 2009 'Big Bang' the value of these coupons was a fair spread (also called par spread), i.e. a spread such that the protection-buyer and the protection-seller legs match at inception. After the 2009 'Big Bang', it is typically a standard contractual coupon of 100bps or 500bps (depending on the credit quality) that is quarterly paid until maturity. An upfront payment, i.e. an amount to be exchanged immediately upon entering the contract, is made between the two counterparties to make the deal fair; Right: In case of 'default', the protection-buyer stops paying coupons to the protection-seller and received from the protection-seller an amount equivalent to the CDS notional minus the recovery; illustrations taken from Wikipedia 'Credit default swap'

These present values depend on interest rates (for discounting the cash flows), but more essentially on the probability of default to know how much cash flows to expect. Pricing a CDS is the conversion of the CDS default probabilities to CDS prices/upfront/spreads. In practice, most practitioners use market-based pricing, i.e. they use quotes (upfront/spread) from the CDS market on some standard and liquid maturities (typically 1, 2, 3, 4, 5, 7, 10-year contracts) to calibrate a model of default probabilities, and then recover the CDS prices (upfront/spread) for any (not necessarily quoted) maturities (for example 3.5 years).

Technically, it is assumed that default is a Poisson process, with an intensity (or hazard rate) $\lambda(t)$. If we note τ the default time, then the probability of default over an infinitesimal time period dt, given no default to time t is $\mathbb{P}(t < \tau < t + dt \mid \tau > t) = \lambda(t)dt$. The probability of surviving to at least time $T > t$ assuming that the default did not occur before time t is $Q(t,T) = \mathbb{P}(\tau > T \mid \tau > t) = e^{-\int_t^T \lambda(s)ds}$. A well-spread assumption is that $t \mapsto \lambda(t)$ is piecewise constant [16, 227] between the quoted maturities as depicted in Figure 3.3. Calibration of the model, i.e. finding default probabilities that are consistent with market prices, is thus easy: simple integration and use of the bootstrap technique [159].

Another pricing tool widely used by market participants is the ISDA standard CDS model `www.cdsmodel.com`. The ISDA standard CDS model is used to convert upfronts to running conventional spreads (given an hypothetic recovery rate, the contractual coupon and interest rates) and conversely. This model has documented shortcomings [227, 16]: it is essentially a conversion tool that yields a semblance of the previous fair spreads (aka par spreads) and as

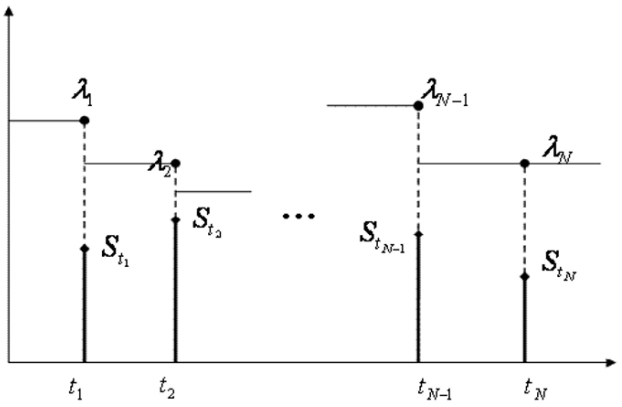


Figure 3.3: Historically, most market participants have used a piecewise constant intensities model to calibrate spread curves from market quotes on 1, 2, 3, 4, 5, 7, 10-year credit default swap spreads.

such acts as a *lingua franca* between the market participants which were accustomed to the more convenient (but contracts are then less fungible) par spreads which are not notional dependent unlike upfronts. These conventional spreads computed by the ISDA model should not be used for modeling purpose, for example one should not strip hazard rates or calibrate models across them [16].

3.1.3 Market structure and participants

There is no exchange or market place to trade credit default swaps. As of today single-name credit default swaps are still traded over-the-counter with market makers in investment banks. The role of these market-makers is to bring liquidity by buying and selling these products and trying to capture the bid/ask spread to remunerate themselves for this service. They work in two different and separated markets (cf. Figure 3.4):

- the inter-dealer market (aka the street) where they trade between themselves using brokers' intermediation to manage their inventories;
- with clients (hedge funds such as Hellebore Capital and other asset management companies) where they act as liquidity providers.

There are about 13 main dealers for credit default swaps: JPMorgan Chase & Co., Goldman Sachs, Morgan Stanley, BNP Paribas, Société générale, Bank of America, Barclays, Citi, Deutsche Bank, Credit Suisse, UBS, HSBC, Nomura. As we will see in section 3.2, these traders send messages containing prices (but not exclusively, it may also contain other information such as market comments) to their clients. Their contribution is unequal and some market makers contribute much more than the others (cf. subsection 3.2.2).

3.1.4 Resources

We conclude this section by listing below a few resources to understand more thoroughly the credit default swap market:

- Data:
  - DataGrapple `www.datagrapple.com`
  - OTCStreaming `www.otcstreaming.com`
  - ISDA `http://www.swapsinfo.org/`
  - Bloomberg (Enter > CDWS)
  - Quandl/Cambridge `https://www.quandl.com/databases/CCDS`
  - Markit `http://www.markit.com/Product/Pricing-Data-CDS`

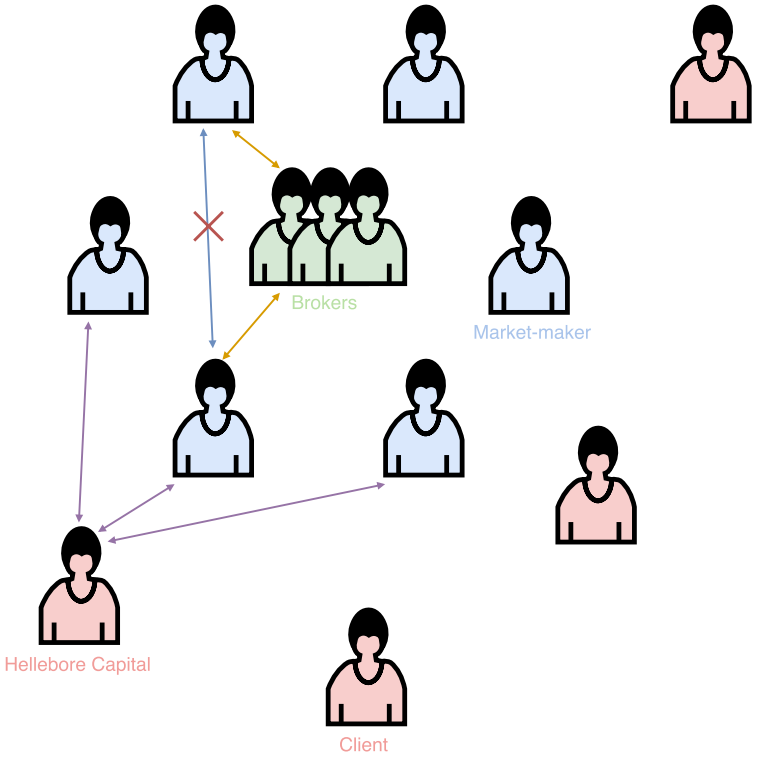


Figure 3.4: Market-makers provide liquidity to their clients by selling and buying them credit default swaps. If needed they can trade between themselves by using brokers services to avoid collusion.

  - Ice `https://www.theice.com/market-data/pricing-and-evaluations/cds_pricing`
  - Thomson Reuters Datastream
    `http://financial.thomsonreuters.com/en/products/tools-applications/trading-investment-tools/datastream-macroeconomic-analysis.html`

- Official documentation:
  - ISDA `http://www2.isda.org/` and more particularly `https://www2.isda.org/asset-classes/credit-derivatives/`
  - Big Bang Protocol (April 8, 2009): `http://www.isda.org/bigbangprot/docs/Big-Bang-Protocol.pdf`
  - Small Bang Protocol (July 27, 2009): `http://www.isda.org/smallbang/`
  - 2014 ISDA Credit Derivatives Definitions `http://www2.isda.org/asset-classes/credit-derivatives/2014-isda-credit-derivatives-definitions/`
  - Single-name CDS Roll `http://www2.isda.org/asset-classes/credit-derivatives/single-name-cds-roll`, December 21, 2015.
  - ISDA Bookstore
    `http://www.isda.org/publications/isdacredit-deri-def-sup-comm.aspx#isdacrd`

- Books:
  - Modelling single-name and multi-name credit derivatives [159]
  - The credit default swap basis [53]
  - An introduction to credit derivatives [52]
  - Interest-Rate Models: Theory and Practice (Chapters on CDS, CDS calibration, CDS options, CDS market models) [39]

  - Credit Models and the Crisis: A journey into CDOs, Copulas, Correlations and Dynamic Models [44]
  - Counterparty Credit Risk, Collateral and Funding, with Pricing cases for all asset classes [41]

- Standard models:
  - ISDA/Markit `http://www.cdsmodel.com/cdsmodel/`

- ISDA research:
  Single-name Credit Default Swaps: A Review of the Empirical Academic Literature [56]
  `http://www2.isda.org/attachment/ODcwMw==/Single-Name%20CDS%20Literature%20Review%20-%20Culp,%20van%20der%20Merwe%20&%20Staerkle%20-%20ISDA.pdf`

- Sell-side research:
  - CDS Curve Trading Handbook 2008 - Quantitative Credit Strategy [175]
    `http://mhderivativesolutions.com/wp-content/uploads/2014/07/6716-Barclays-Capital-CDS-Curve-Trading-Handbook-20081.pdf`

- Academic research with practical concerns (the following list does not intend to be exhaustive at all):
  - [31], [34], [40], [30], [32], [43], [42], [13], [208], [37], [35], [38], [44], [45], [36], [33], [16], [47], [46]

    Notice that the works [32, 43] also deal with a notion of clustering: the clustering of defaults. Authors (Brigo *et al.*) propose the first arbitrage-free dynamic loss model that can be calibrated cross-sectionally and consistently to tranches (of a credit derivatives portfolio) for different attachment and detachment points and maturities. They observe that clusters (modes) appear on the tail of the implied loss distribution coming from the model calibration under the pricing measure. This clustering in the loss distribution is analyzed through the financial crisis in [38], and the youtube video `https://www.youtube.com/watch?v=YZO-HeaGHkk&t=62m40s` shows an animation with this clustering behaviour through the crisis.

    Though interesting for credit derivatives risk management, this type of clustering (implied clustering of defaults with respect to the pricing measure) is not investigated further in this thesis. It could have been a fruitful avenue of research to compare the implied clustering (of defaults) versus the historical clustering based on a tail-dependence measure: Does the hierarchical clustering structure based on a tail-dependence measure estimated on historical data become sharper when the clusters in the implied distribution loss appear? The question is left open.

3.2 A database of CDS quotes sent by dealers

The market structure for credit derivatives, and in particular for credit default swaps and credit default swaps indices, keeps transforming and modernizing. In its "Big Bang" of April 2009, the International Swaps and Derivatives Association (ISDA) introduced a number of documentation changes where the single-name CDS contracts have been standardized (`http://www.isda.org/press/press040809.html`). After that major standardization phase, as one of the consequences of the 'Dodd-Frank Wall Street Reform and Consumer Protection Act' vote, transaction mechanisms for CDS indices have evolved: CDS indices can be (but not exclusively) traded on Swap Execution Facilities (SEFs), which are electronic platforms for financial swaps trading. SEFs provide pre-trade information such as bid and offer quotes. SEFs also record the trades which are done on the platforms and they have to release publicly part of the information in swap default repositories (SDRs) in order to comply with the

'Dodd-Frank Act' (cf. Section 3.3). However, as of today, messages (such as emails or Bloomberg chats) remain the mainstream way to convey information between dealers and their clients.

This section presents Hellebore Capital's proprietary database of credit derivatives messages. We first describe in subsection 3.2.1 a few typical messages and explain their content. Then, we propose a descriptive study of the database volumetry and its main characteristics in subsection 3.2.2. We also briefly describe in subsection 3.2.3 how we convert this raw data into time series.

3.2.1 Zoology of credit default swaps messages

The ever growing database of messages is as of today the main source of information for Hellebore Capital upon which models, decision making and other processings (market monitoring, valuation, risk calculations) rely on.

Hellebore Capital receives about 20,000 messages a day which add to the database counting more than 40,000,000 messages in total.

Messages contain bid and ask prices for CDS, CDS indices, options on CDS indices, tranches, bonds, loans and other bespoke products. Some may contain market comments from traders or simply financial news that were forwarded to Hellebore's mailbox. We display in Figures 3.5, 3.6, 3.7, 3.8, 3.9, 3.10, 3.11, 3.12 a few typical examples of messages received from Hellebore Capital's counterparties, i.e. the main credit dealers, to illustrate the raw material.

The message depicted in Figure 3.5 is a typical example of a CDS 'run': it contains a list of bids and asks (which are separated by the "/" character) for a subset of credit default swaps. The trader who had sent it is in charge of the 'Automotive' sector and showed Hellebore Capital quotes on the 5-year credit default swaps on these companies (BMW, Daimler, Ford Credit Europe, Fiat, Peugeot, Porsche, Renault, Jaguar and Volkswagen). Quoted entities are designated here by both a 'long name' (rightmost column in the message) and a 'ticker' (leftmost column). Most of the time, only tickers are used. There may exist several 'reference' ticker (from Bloomberg, Markit, etc.) for designating a given credit default swap. Their use is at the counterparty discretion. The market-maker also gave us 'curve switches' (columns 3s5 and 5s7), i.e. the price difference between a 3-year contract and a 5-year contract (3s5), here 15/21 for BMW, and the price difference between a 5-year contract and a 7-year contract (5s7), here 12/20 for BMW. We can notice that the price for a 2 year extension of the contract between a 3-year maturity to a 5-year one is more expensive than the price for extending from a 5-year maturity to a 7-year one. The CDS term structure is concave which is the standard behaviour for unstressed entities. We can also notice that the market-maker did not give the 'switch' information for all the single-names (FIAT, PEUGOT, TTMTIN curves are missing). In the third column from the left, the market-maker indicated some additional information: the price change since the previous end of day.

The message displayed in Figure 3.6 is about some European 'High-Yield' single-name credit default swaps. Unlike the previous one which is focused on a particular sector (automotive industry), this one gathers entities having a 'High-Yield' (non investment grade) rating and which are members of the iTraxx Crossover (XO) index. The market-maker indicated that this information is conditional to a given price of the index: "XO = 239.50". In this message, there are only information for the 5-year credit default swaps (not the curves), but more detailed:

- Both a ticker (leftmost column) and a longer name (rightmost column),
- Bid and ask spreads separated by "...",
- Indicative notional sizes (in million of euros) for which these quotes are valid: The market-maker gave the size for both the bid and ask spreads separated by the character "x". Notice that they are the same but for one entity (Hellenic Telecom, 2x0) which incurred a strong widening of its spread: +10bps (indicated in the 'CoD' column).

FROM:TraderX (DEALER1)
[SENT]2014/12/01 17:01:16 [LOCAL]2014/12/01 18:01:16 [NonForwardable]
DEALER1 Euro Autos CDS 5 5:00PM

	5yr		3s5	5s7	
BMW	35/38	-	15/21	12/20	BMW
DAIGR	36/39	-	16/22	13/21	Daimler AG
FCE	70/85	-	32/48	25/40	FCE Bank PLC
FIAT	215/227	+4			FIAT SPA - FCA
PEUGOT	225/237	+5			Peugeot SA
PORSCHE	40/55	i	15/25	10/25	Porsche
RENAUL	93/100	+0.5	38/48	30/45	Renault SA
TTMTIN	122/134	-			Jaguar Land Rover JAGLN
VW	44/47	-	16/22	13/21	Volkswagen AG

Figure 3.5: A CDS run on the automotive sector

- The CDS standard coupon is indicated (typically 100 or 500 bps). This information is mandatory to correctly interpret the CDS spread, but is often left implicit in the messages.
- A market on recovery rate swaps (column "RR Swap"). They are another kind of derivatives whose purpose is to hedge the recovery risk in case of a default. They essentially become traded when an entity is near default, otherwise they take a standard and conventional value of 40% for European credit default swaps referencing senior debt (20% for subordinated debt; other standard rates for Asian and American entities).

BankY HY: TMT/ Auto 5yr CDS Markets XO = 239.50 #5

			Str	CoD	RR Swap	
Altice	425...445	2x2	500	0	35..45	Altice Finco
Alufp	362...382	2x2	500	2	35..45	Alcatel
Atcna	351...381	QxQ	500	6	35..45	Altice Sa
Certch	266...296	2x2	500	1	35..45	Cerved
Cirim	75...95	2x2	500	0	35..45	CIR Spa
Cwcln	270...290	2x2	500	0	35..45	Cable & Wireless
Fiat	240...260	3x3	500	0	35..45	Fiat
Havfp	65...85	2x2	100	1	35..45	Havas
Htoga	365...395	2x0	500	10	35..45	Hellenic Telecom
Itvln	117...137	5x5	500	5	35..45	ITV Plc
Nokia	117...137	5x5	500	0	35..45	Nokia
Numfp	322...342	5x5	500	-3	35..45	Numericable
Nxp	145...165	3x3	500	-2	35..45	NXPBV
Peugot	250...270	5x5	500	0	35..45	Peugeot

Figure 3.6: A run on some European "High-Yield" CDSs, mostly members of the XO index

The two previous messages were describing bid and ask prices for a particular subset of credit default swaps quoted by a given trader. A market-maker may sometimes be particularly interested to clean positions from his portfolio and can indicate it very clearly such as in the message depicted in Figure 3.7. We can learn from it that this trader is willing to buy some names (e.g., "BUYER ABIBB") and sell others (e.g., "SELLER BERTEL") at the proposed levels (indicated by the character '@' followed by the spread in bps; here it is implicit that the focus is on 5-year maturity contracts with standard coupon of 100 bps). Since market-makers at the moment of sending axes are generally willing to trade and reveal their interest, they propose to their clients rather aggressive prices (at least according to their market view) to win the trades from their competitors (other market-makers in the major investment banks). Notice that this information was also somewhat present in the previous message (in Figure 3.6): proposed sizes for Hellenic Telecom were 2x0 meaning that the market-maker did not want to offer the CDS.

The message in Figure 3.8 was sent by a "High-Yield" trader. It gives several interesting pieces of information besides the 5-year prices such as the standard coupons which are implicitly 500 bps but for one CDS (Ladbrokes PLC) which has a coupon of 100 bps

```
FROM:TraderX (BankY)
[SENT]2014/11/17 11:37:52  [LOCAL]2014/11/17 12:37:52  [NonForwardable]
BankY TMT/RETAIL CDS AXES REFRESHED...
BUYER  ABIBB  @  57     SELLER PSON   @  58
SELLER BATSLN @  46     SELLER RIFP   @  65
SELLER BERTEL @  41     SELLER SESGFP @  59
BUYER  BNFP   @  48     BUYER  STM    @ 118
BUYER  CARLB  @  90     BUYER  SZUGR  @ 128
SELLER JTI    @  29     BUYER  TELEFO @  84
SELLER KPN    @  38     BUYER  TSCOLN @ 139
SELLER MCFP   @  38     SELLER UNANA  @  27
SELLER MKS    @ 114     BUYER  VIVFP  @  66
BUYER  MRWLN  @ 161
BUYER  PHG    @  48
BUYER  PORTEL @ 293
```

Figure 3.7: Axes for investment grade 'retail', 'media' and 'telecommunications' single names

(indicated by the character "#" explained in the message header), price changes and alternative tickers for some companies (e.g., BAB and IAGLN for British Airways PLC; UNITY, LBTYA, IESY for Unitymedia KabelBW GmbH; NXPBV, NXPI for NXP B.V.). Quotes for Air France-KLM, Altice SA, Eileme 2 AB, Scandinavian Airlines System Denmark-Norway-Sweden, TVN Finance Corp III AB, Vougeot Bidco PLC are only 'indicative' (cf. character "*" in the message header). The trader is showing us that he is not willing to trade at these levels or on these names at the moment.

```
From: Delaer2 (Bank X)
Sent: 17 November 2014 08:56:54 (UTC+01:00) Brussels, Copenhagen, Madrid, Paris
Subject: Bank X HY:  +++   HY CDS  5y   OPENING   +++

XO S22 ref     360/353  +5    [* indic, # cpn=100, + sub]

AFFP      5y   115/130  +0    Societe Air France
AFKLM     5y   440/490  -- *  Air France-KLM
ALTICE    5y   420/440  +5    Altice Finco SA
ATCNA     5y   365/395  +5 *  Altice SA
ALUFP     5y   367/382  +5    Alcatel Lucent
CWLN Ltd  5y   270/285  +3    Cable & Wireless Limited
HTOGA     5y   355/385 +10    Hellenic Telecommunications Organization SA
IAGLN     5y   147/167  +2    British Airways PLC [BAB]
IESY      5y   203/218  +3    Unitymedia KabelBW GmbH [UNITY, LBTYA]
LADLN     5y   300/320  +5 #  Ladbrokes PLC
NOKIA     5y   119/129  +1    Nokia OYJ
NUMFP     5y   338/348  +5    Numericable Groupe SA
NXPI      5y   150/165  +5    NXP B.V. [NXPBV]
ONOSM     5y    30/ 50  +0    Ono Finance II PLC
POLKOM    5y   105/125  +0 *  Eileme 2 AB
PFOURS    5y   185/200  +3    Play Finance 1 SA
SAS       5y     4/6   --- *  Scandinavian Airlines System Denmark-Norway-Sweden
SOLSM     5y   340/355  +3    Melia Hotels International, S.A.
SUNCOM    5y   200/215  +3    Sunrise Communications Holdings S.A.
TUI       5y   212/227  +2    TUI AG
TVNPW     5y   155/175  +0 *  TVN Finance Corp III AB
UPC       5y   235/250  +3    UPC Holding B.V.
VMED      5y   215/230  +3    Virgin Media Finance PLC
VUECIN    5y   430/460  +5 *  Vougeot Bidco PLC
WINDIM    5y   437/457  +5    Wind Acquisition Finance S.A.
```

Figure 3.8: A CDS run on some European "High-Yield" single names including indicative quotes

The message displayed in Figure 3.9 has a rich content as it describes several different contracts for a given entity:

- Bids and asks spreads for the most quoted maturities, i.e. 1, 3, 5, 7, 10-year contracts;
- Bids and asks spreads (also the recovery rate in column 'REC') for the two different ISDA definitions: 2003 and 2014 (cf. the 'ISDA' column).

The trader also gave the price to switch from the old 2003 definition to the new 2014 credit default swaps revised definition updated by the ISDA. We can notice that buying protection is more expensive (higher spread) for the new ISDA 2014 definition than for the ISDA 2003 one. Indeed, the chances of triggering the CDS in case of a default are slightly higher due to a wider definition of what constitutes a default. These amendments were incentivised by the experience of the great financial crisis of 2007-2008 and the Greek default to better tackle government bail-ins of banks and sovereign credit event mechanism.

SOVS: CORE

CORE	ISDA	REC	1Y	3Y	5Y	7Y	10Y
AUSTRIA	2003	40	0.5/4.5	7/11	17/21	25.5/32.5	37/45
AUSTRIA	2014	37	-0.4/5.6	6.5/12.5	17/23	26/35	38.1/48.1
AUSTRIA	SWITCH	3	0/1.1	0/1.5	0/2	0.5/2.5	1.1/3.1
BELGIUM	2003	40	8/14	24/30	42/46	58/64	78/86
BELGIUM	2014	36	7.8/15.8	24.4/33.4	43.1/51.1	59.3/69.3	80.9/92.9
BELGIUM	SWITCH	4	0/1.8	0.4/3.4	1.1/5.1	2.3/6.3	3.9/7.9
GERMANY	2003	40	0.5/3.5	5.5/8.5	14/16	22/26	31.5/37.5
GERMANY	2014	37	0.1/4.1	4.9/9.9	13.8/17.8	21.2/27.2	31.3/39.3
GERMANY	SWITCH	3	0/0.6	0/1.4	0/1.8	0.2/2.2	0.8/2.8
NETHERLANDS	2003	40	0.5/4.5	7.5/11.5	17/21	26.5/32.5	37.5/45.5
NETHERLANDS	2014	37	-0.4/5.6	6.5/13.5	16.5/23.5	26/35	36.6/47.6
NETHERLANDS	SWITCH	3	0/1.1	0/2	0/2.5	0/3	0.6/3.6
UK	2003	0	0.5/4.5	6/10	15.5/17.5	23.5/27.5	33/40
UK	2014	40	0.2/5.2	5.6/11.6	15.7/19.7	23.3/29.3	33.6/42.6
UK	SWITCH	-40	0/0.7	0/1.6	0.2/2.2	0.8/2.8	1.6/3.6

Figure 3.9: A full CDS run with curves for European sovereigns (2003 and 2014 ISDA definitions)

Close to the roll dates (March 20 and September 20 of each year since December 21, 2015 - before this date single-name credit default swaps were rolled quarterly: every March 20, June 20, September 20, December 20) market-makers send messages such as those displayed in Figure 3.10. They indicate how much the client has to pay to roll its current 5-year CDS (which is going to be *off-the-run* after the roll date, i.e. much less liquid since its 4.5-year equivalent maturity is not standard and quoted) into a 5.5-year contract (as indicated by 'ROLLS 5-5.5y'). The 5.5-year contract will become a plain 5-year CDS after the roll date and as such will *de facto* benefit from the best liquidity possible in the market. In Figure 3.10, the message on the left only indicates rolls for Latin American sovereign credit default swaps, but not the outright levels; the message displayed on the right gives both outright 5-year prices and the '5-5.5y' rolls for US insurance credit default swaps. Notice that the price for a 6-month extension of a 5-year contract corresponds roughly to 1/10th of the price for the 5-year contract; only roughly since the term structure is not necessarily linear.

The message in Figure 3.11 is typical of a market on indices: a one-liner in the subject of the email, repeated in the body, and containing index vintage (usually the *on-the-run* one), bid and ask spreads, index ticker. Here, the iTraxx Crossover ("XO", in short) and the iTraxx Europe (aka the "Main") are the indices concerned. The maturity is implicitly 5 years since index vintage "S22" was the *on-the-run* one at that time. Sometimes this information is also implicit. This kind of message is sent with a relatively high-frequency by the dealers (several times an hour) as these indices are particularly liquid.

At a much lower frequency, market-makers also send messages containing prices for other less liquid maturities (such as the 3, 7, 10-year in Figure 3.12) and *off-the-run* index vintages (such as IG22, IG21, IG20, ..., IG4 in the message displayed in Figure 3.12).

LatAm CDS Roll	
Brazil	20.75/21.00
Mexico	15.00/15.50
Colomb	16.5/17.25
Peru	13.25/13.75
Arg	31.00/34

TRADING INSURANCE
OUTRIGHT 5y
PRU 59/63
MET 73/77
AIG 78.5/82.5
HIG 45/49

ROLLS 5-5.5y
PRU 9.5 / 12.5
MET 10.5 / 13.5
AIG 8.5 / 11.5
HIG 5 / 8

Figure 3.10: Latin American sovereign rolls (left); 5-year spreads (top) + rolls (bottom) for US insurance credit default swaps (right)

From: Trader2 (BankY,)
Sent: 24 November 2014 16:05:52 (UTC+01:00) Brussels, Copenhagen, Madrid, Paris
Subject: BankY - S22 XO 338.5-340 [-7.75], MAIN 59.375-59.625 [-1.625]

BankY - S22 XO 338.5-340 [-7.75], MAIN 59.375-59.625 [-1.625]

Figure 3.11: A market on two European CDS indices: iTraxx Europe and iTraxx Crossover

	5y	3y	7y	10y
IG23	60¾ref	32/35	81/84½	101¼/102¾
IG22	52⅜/53⅜	26¼/29¼	76⅝/78⅝	97⅛/99⅛
IG21	44⅜/45⅜	17⅝/21⅝	70⅛/73⅛	92⅛/94⅛
IG20	37⅜/38⅝	13½/17½	63½/66½	89⅜/91⅝
IG19	30¼/31¾	6¼/11¼	57/60	86/88½
IG18	24¼/26¼	4/10	49⅜/52⅜	81¼/84¼
IG17	17/19½	1¾/8¾	41/44½	77/80½
IG16	13½/16		35/38½	72½/76
IG15	7¾/10¾		27½/31½	67¼/71¼
IG14	5/9		22¾/27¼	62¼/66¼
IG13	2½/8½		15¾/20¼	53¾/57¾
IG12			14½/18½	51/55
IG11			12¾/17¾	52¼/57¼
IG10			10¾/16¾	49¾/54¾
IG9			11½/16½	48¼/50¼
IG8				42¾/47¾
IG7				91½/97½
IG6				104¼/112¼
IG5				123¼/133¼
IG4				132¾/147¾

Figure 3.12: A run on CDX North American Investment Grade index: curves and *off-the-run* vintages

3.2.2 Descriptive statistics of the database

In this subsection, we briefly describe the content of Hellebore Capital's database in terms of volumetry through the graphs displayed in Figures 3.13, 3.14, 3.15, 3.16, 3.17, 3.18, 3.19, 3.20, 3.21, 3.22 (produced by Philippe Very).

In Figure 3.13, we display the number of quotes received daily on CDS and CDS indices. We can see that these numbers peaked during the European debt crisis (about 80,000 quotes a day).

If we consider only 5-year credit default swaps which are by far the most liquid ones, we can see in Figure 3.14 that they amount for about half of the total number of quotes received each day.

In Figure 3.15, we focus on the daily number of 5-year quotes received by geographical region (Asia, US, Europe) subdivided by rating (investment grade or high-yield). We can see that the most contributed 5-year credit default swaps are those on the Europe and US investment grade entities. Relatively few quotes are available on Asia credit default swaps.

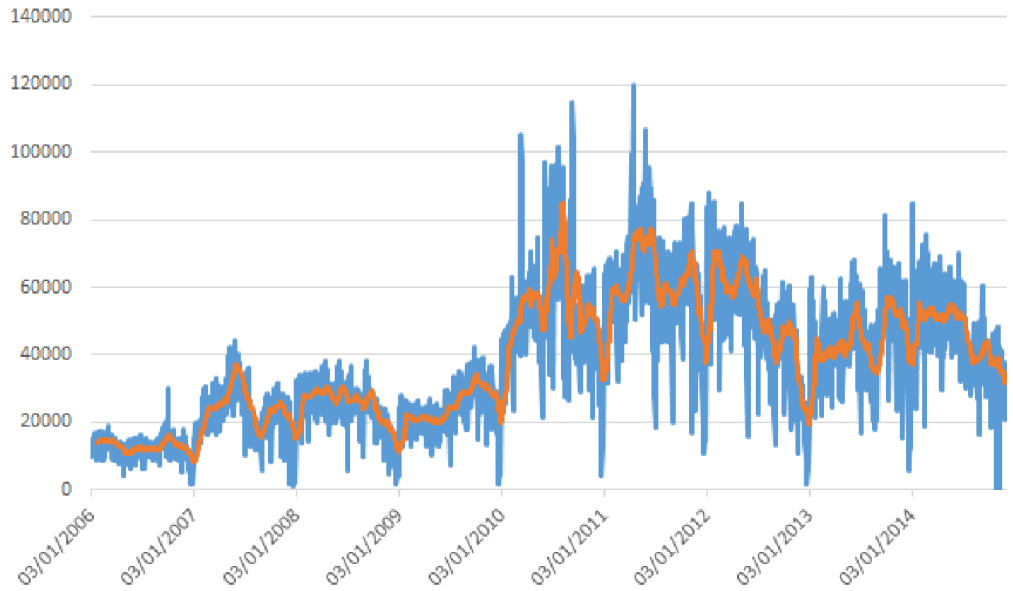


Figure 3.13: Daily number of quotes received concerning CDS and CDS indices

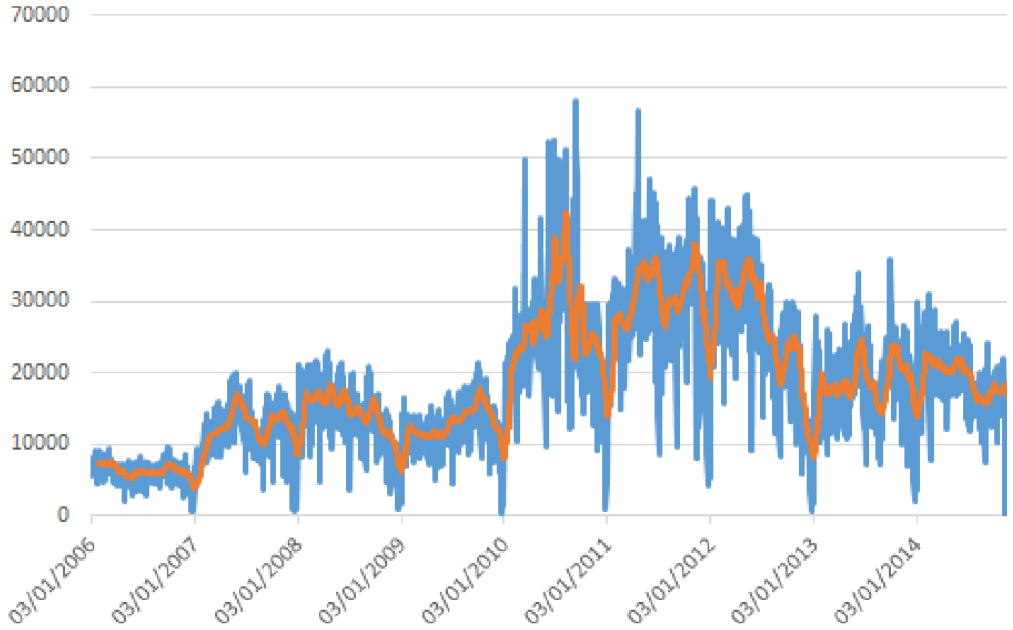


Figure 3.14: Half of the quotes received daily relate to 5-year credit default swaps

In Figure 3.16, we display the proportion of 5-year quotes received from the different dealers by 'region - rating'. We can see that for Asian High-Yield entities more than 60% of the 5-year quotes were sent by a single dealer (Dealer 2). This dealer is also very active in sending information on sovereign 5-year credit default swaps. For Europe and US, investment grade and high-year, Dealer 1 is a major contributor. He sends 50% of the total number of quotes on US entities, 35% of the total number of quotes. Dealer 2 and 3 have sent respectively 13% and 12% of the total number of quotes for 5-year credit default swaps.

In Figure 3.17, we display the number of quotes received each hour during an average trading day. Notice that we receive some quotes during the night which correspond to quotes on Asian entities; They are sent between 23pm until the Asian market is closed at roughly 10am. The bulk of quotes is received between 6am and 8pm. We receive the highest number of quotes during the day when both the European and US markets are open, i.e. between 11am and 16pm. For each market (Europe and US), a peak is observed at the start of the trading day (7am and 12am respectively) as market-makers may send full runs with the opening prices.

In Figure 3.18, we display the proportion of the quotes received for different CDS ma-

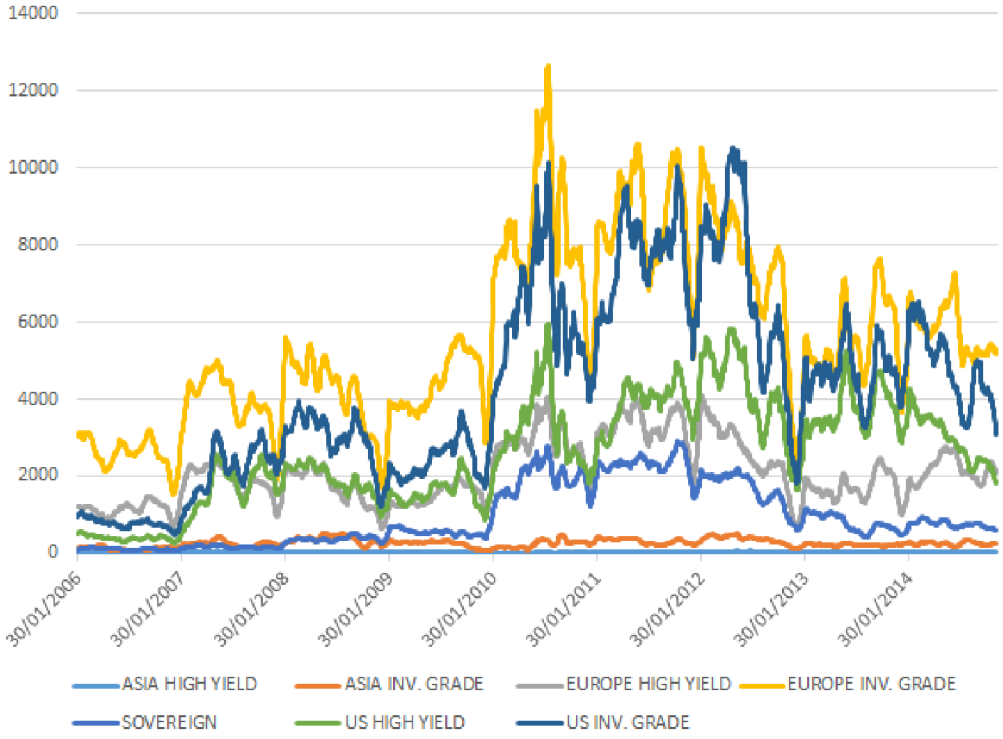


Figure 3.15: Daily number of quotes for 5-year credit default swaps by 'region - rating'

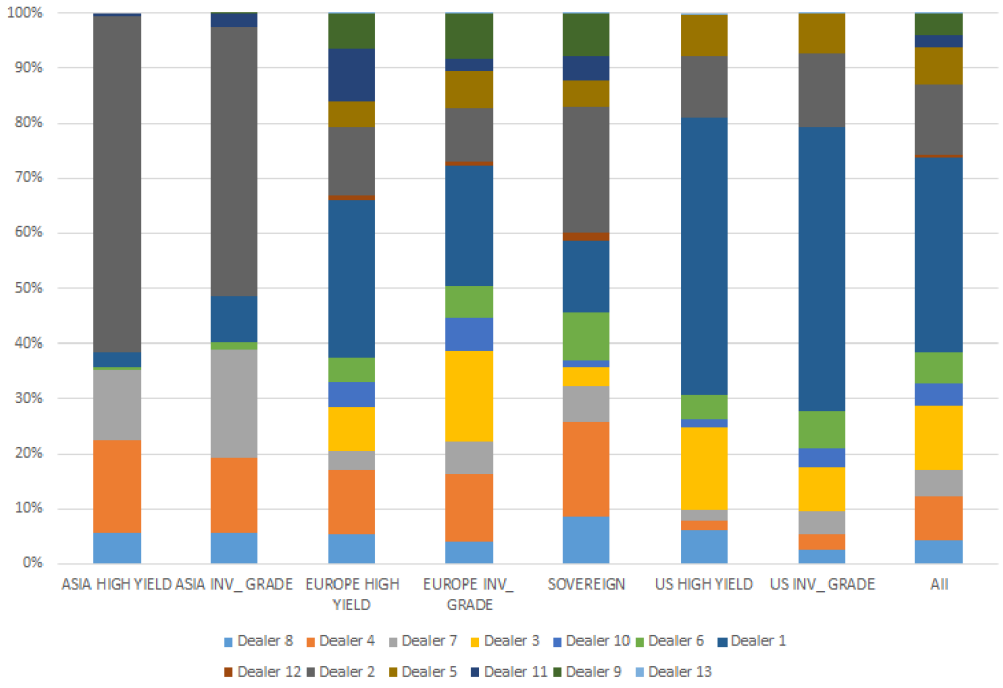


Figure 3.16: Proportion of 5-year quotes received from the different dealers by 'region - rating'

turities by 'region - rating'. We can notice that in Asia, most of the quotes received relate to 5-year contracts. For Asian investment grade entities, the quotes on 5-year credit default swaps constitute more than 90% of the information received; 60% for European entities and 50% overall. To see better how the quotes are distributed between the less liquid maturities, we display the same graph excluding the 5-year quotes in Figure 3.19.

Overall quotes on maturities other than the 5-year one are rather equally distributed: 20% for the 3,7-year quotes, 14% for the 4,10-year quotes. However, the proportion of 10-year quotes significantly varies across the regions: 13% and 7% for US investment grade and US high-yield respectively, but more than 40% for sovereign entities; almost no 10-year quotes for Asia high-yield entities. For the latter ones, the 3-year CDS maturity is the most

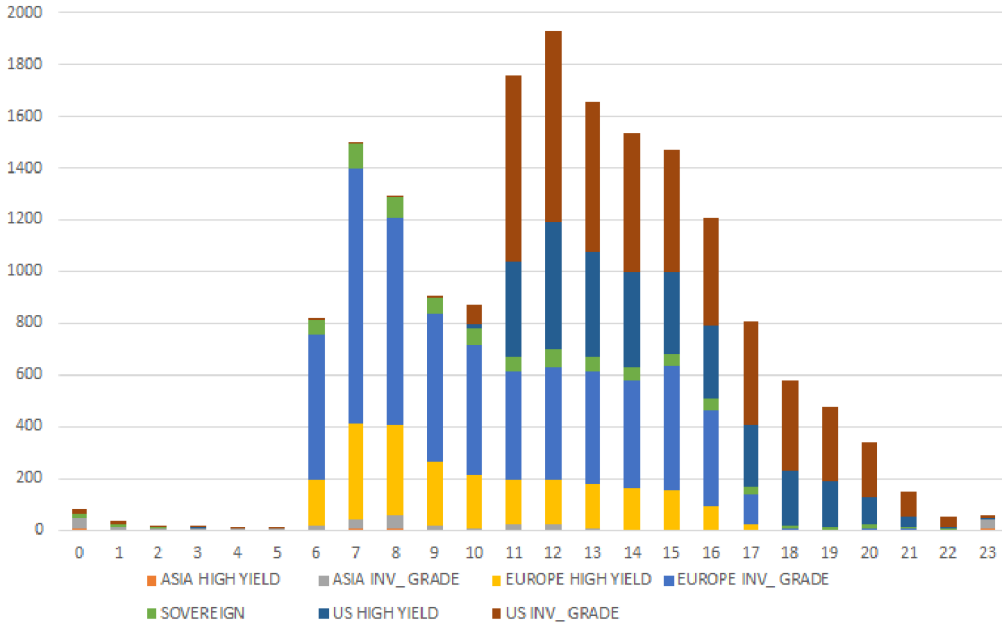


Figure 3.17: Average number of 5-year CDS quotes received each hour during a trading day; In color, the 'region - rating' proportion concerned by the quotes is indicated

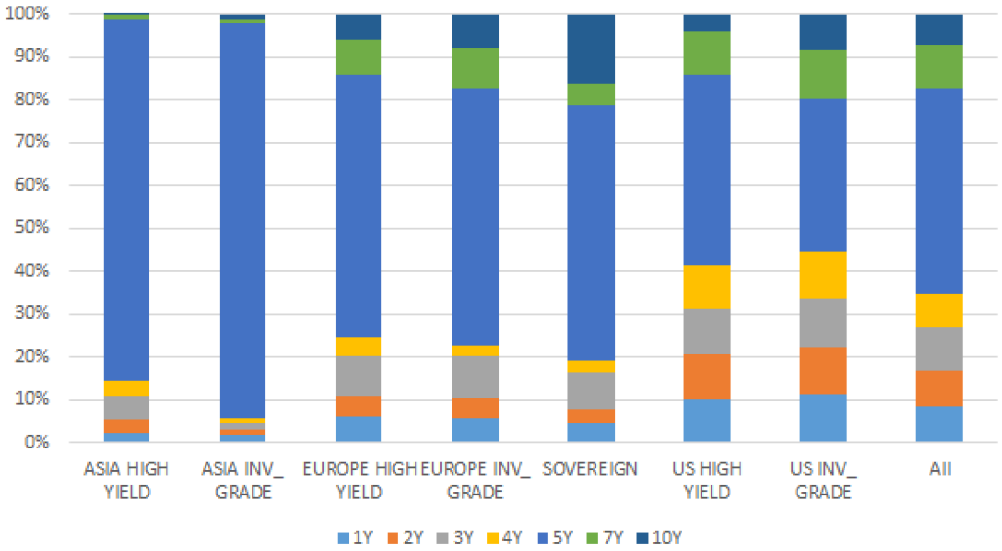


Figure 3.18: Proportion of the quotes received for different CDS maturities by 'region - rating'

significant one (excluding 5-year quotes): 35% of the information on curves.

In Figure 3.20, we display the proportion of the quotes received from the different dealers by CDS maturity. We can notice that most of the information on curves (about 70% overall) was sent from a single dealer (Dealer 1), especially for 1,2,4-year credit default swaps. For US entities, Dealer 1 have sent 80% of all the curves. For European entities, 'only' 50% of the quotes on curves were sent by Dealer 1. Dealer 2, Dealer 4 and Dealer 7 had also a significant contribution with 15%, 16% and 12% of the total number of quotes on European curves respectively. Dealer 1 is also the one who sends most of the information on 5-year contracts (about 35%).

In Figure 3.21, we display the proportion of received quotes on curves (maturities other than the 5-year one) by 'region - rating' and by 'sector'. We can notice that most of them relate to European investment grade entities (27% of the quotes on curves) and US investment grade entities (38%). We do not receive many curves on Asian entities (less than 1%). Concerning the distribution of received curves by sector, 22% of them relate to financial entities and 32% to consumer goods entities (17% cyclical, 15% non-cyclical); for other

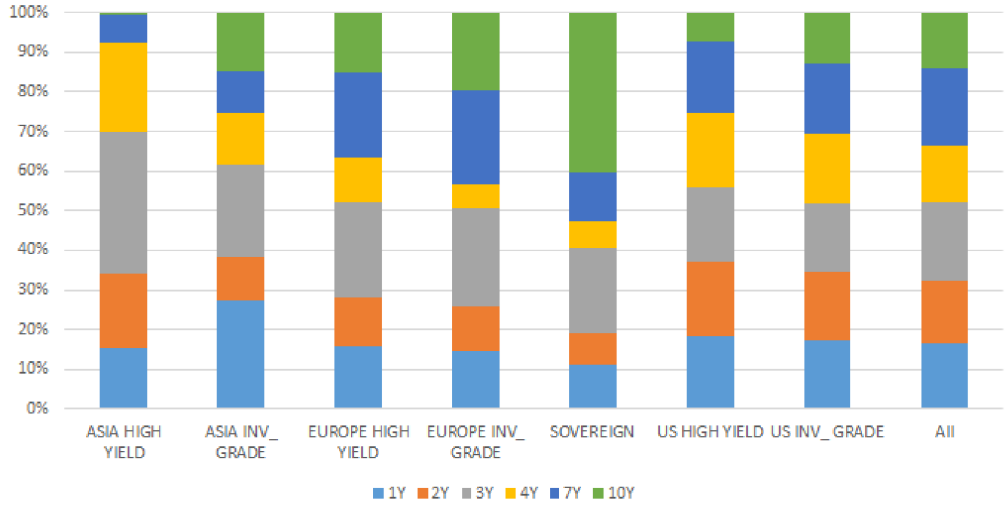


Figure 3.19: Proportion of the quotes received for different CDS maturities (without the most contributed 5-year) by 'region - rating'

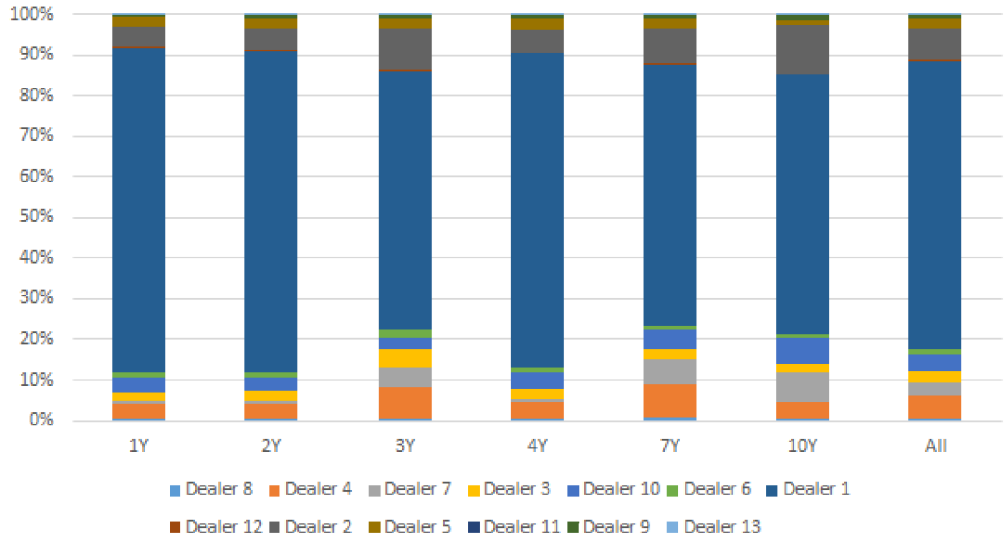


Figure 3.20: Proportion of the quotes received from the different dealers by CDS maturity

sectors, curves are roughly equally distributed (around 8%) but for the technology sector (only 2%).

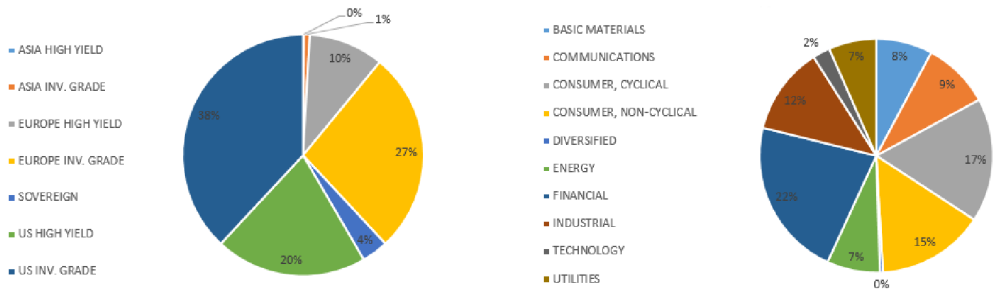


Figure 3.21: Proportion of received CDS curves (1,2,3,4,7,10-year) by 'region - rating' and by 'sector'

In Figure 3.22, we show the distribution of the quotes received from the different dealers by CDS index. Dealer 1 is still the most significant provider of information with 28% of the quotes sent overall followed by Dealer 4 (19%), Dealer 3 (13%) and Dealer 2 (12%). However, their contribution strongly depends on the index family: Dealer 7 (30%) and Dealer 4 (26%) are the major provider of quotes on the US investment grade indices (CDXIG family) whereas

Dealer 1 had sent only 7% of the CDXIG quotes; On financial indices (ITXES and ITXEU), Dealer 1 is the major one by far with 37% of all the quotes sent.

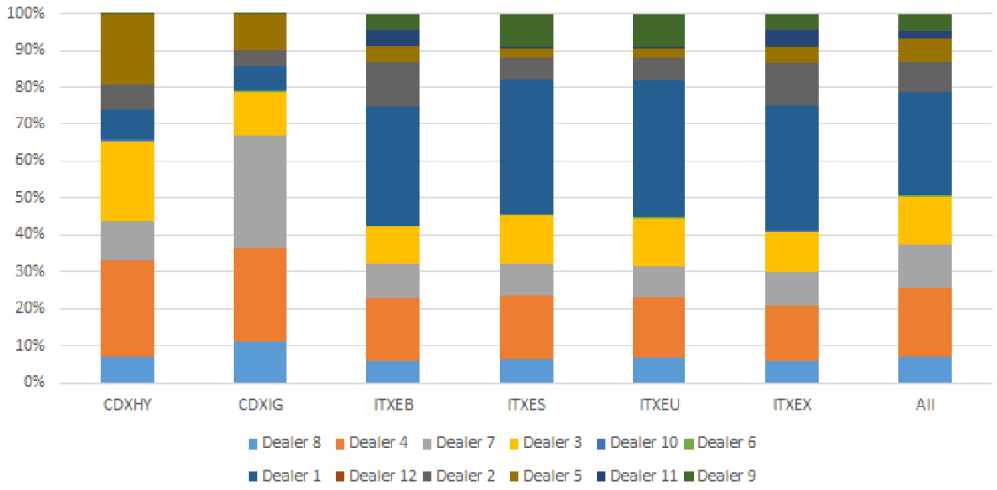


Figure 3.22: Proportion of the quotes received from the different dealers by CDS index

3.2.3 From quotes to time series

One of the service provided by Hellebore Technologies is the accurate extraction of information from these messages into a standard structured database (`www.scriptminer.com`). Part of the methodology using modern Natural Language Processing techniques such as token embeddings is presented in [200] and has been the subject of Marc Szafraniec research internship.

In Figure 3.23, we show the quotes received on the 5-year CDS of Banco Santander, S.A. during one week (between September 6, 2016 and September 14, 2016). We can notice that several market-makers (11 market-makers that have been anonymized as A, B, ..., K) have contributed to quote its spread according to their market views, the flows from clients, the content of their inventory and their risk limits. They send these bid and ask quotes more or less frequently and not necessarily regularly. We thus receive asynchronously these price updates from the different market-makers.

This information is then fed to the many services (market-based valuation of credit derivatives portfolios, risk calculation, market monitoring, backtesting quantitative strategies, etc.) running at Hellebore Capital.

One of the first such processing is the building of synthetic order books. These synthetic order books, which are inspired by the order books on organized markets, help to summarize efficiently all this information. In our synthetic order book, quotes are ordered from the best bid to the worst one, same for ask quotes so that it gives Hellebore Capital the best market conditions (counterparties and spreads) when it decides to trade. We can visualize these real-time 'best bid / best ask' in Figure 3.24. To effectively implement these synthetic order books, two assumptions are made: (i) price persistence (we do not know for how long a quote is valid, it is not necessarily valid until we receive a new one from the same market-maker) up to a certain amount of time; (ii) priority to certain dealers which are known to be more reliable. The thorough study and modeling improvement of these synthetic order books is one of the research axes at Hellebore Capital. Mikołaj Bińkowski, a PhD student supervised by Prof. Rama Cont at Imperial College London, is partially working on this topic [19].

Once the 'best bid / best ask' quotes have been identified on all the products, further processings can be applied to their mid-price (arithmetic average of the best bid and best ask quotes). For example, one can build historical daily time series by taking snapshots of these mid-prices at 5:00 PM GMT (convenient but arbitrary closing time which helps to alleviate many statistical issues due to the different market closing hours) every trading

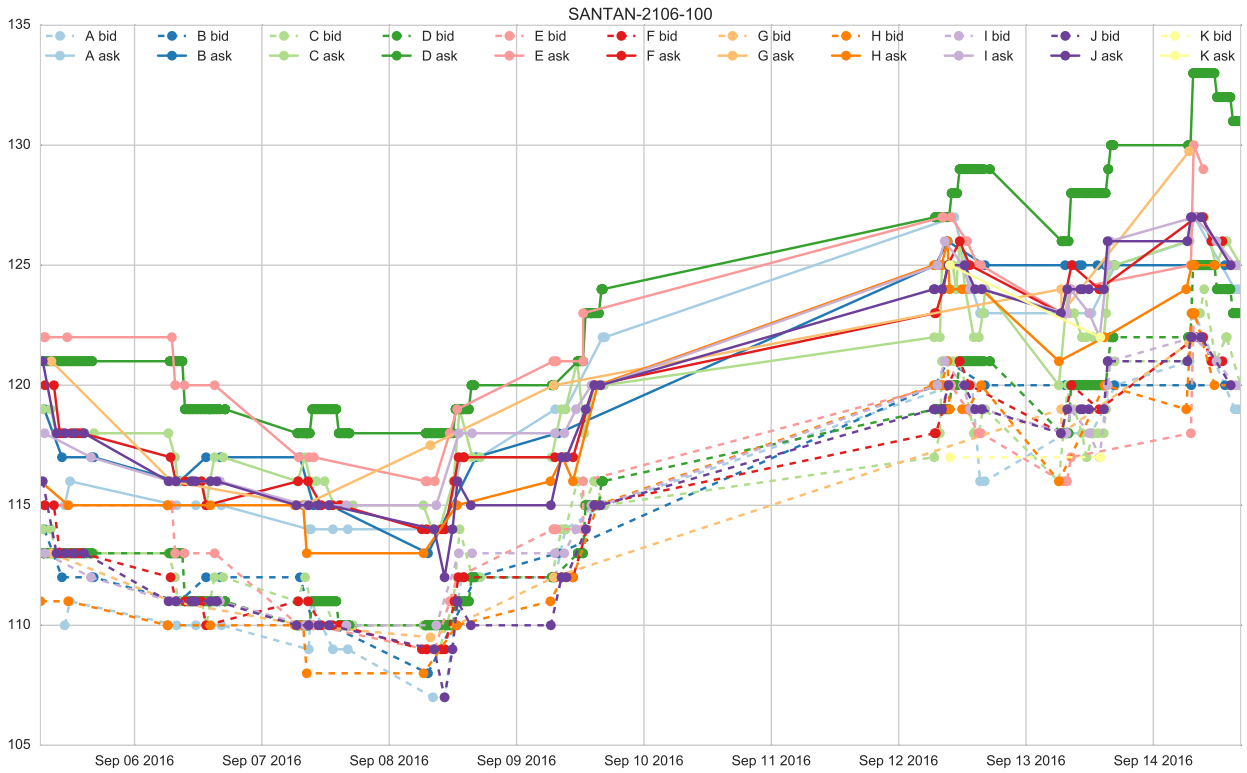


Figure 3.23: Quotes on Santander CDS received from 11 different market-makers during one week

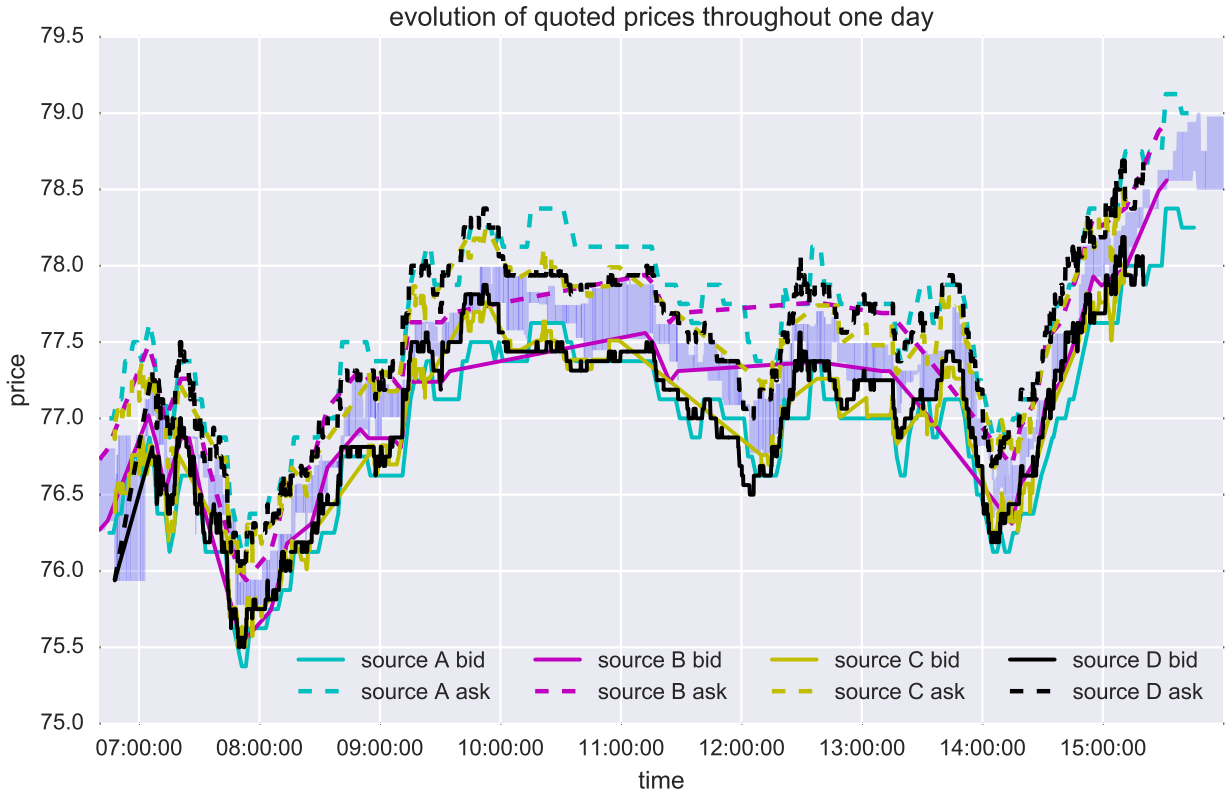


Figure 3.24: Quotes from four different market participants (sources) for the same CDS (iTraxx Europe Main Index, a tradable Credit Default Swap index of 125 investment grade rated European entities) throughout one day. Each trader displays from time to time the prices for which he offers to buy (*bid*) and sell (*ask*) the underlying CDS. The filled area marks the difference between the best sell and buy offers (*spread*) at each time.

days. It yields time series such as those displayed in Figure 3.25 and available in DataGrapple (`www.datagrapple.com`) starting January 2006 and still growing as of today...

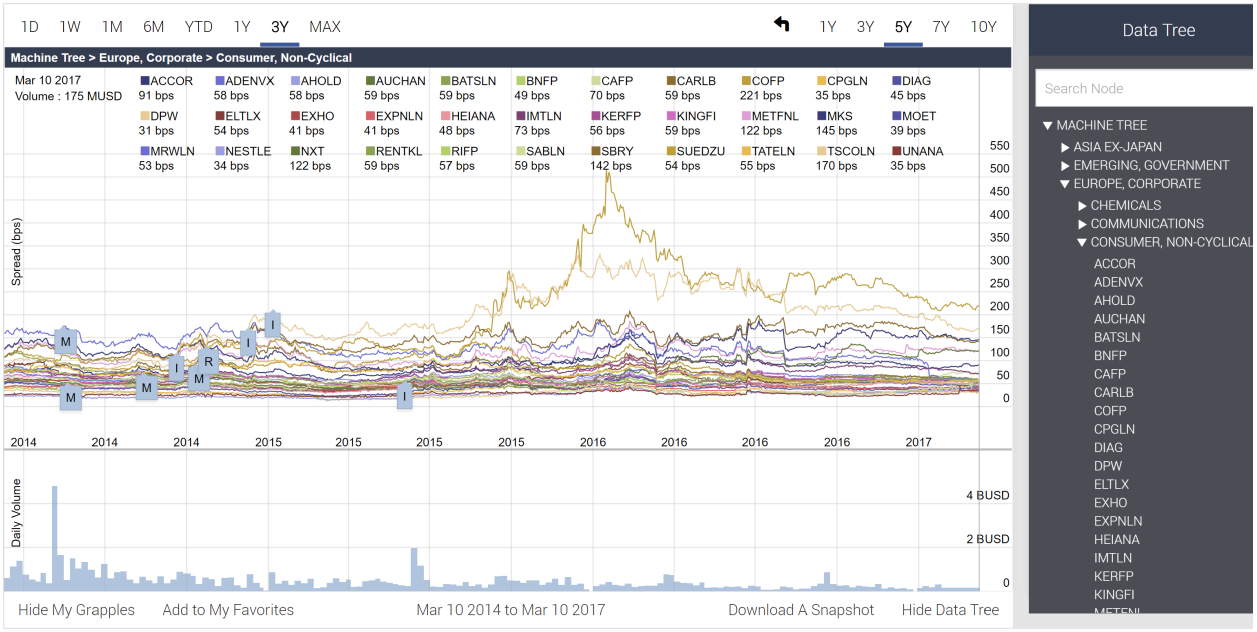


Figure 3.25: Historical time series of 'Consumer, non-cyclical' 5-year CDS spreads (with coupon 100) over the last 3 years; DataGrapple (`www.datagrapple.com`), powered by Hellebore Technologies, is a web portal that allows to browse efficiently thousands of time series using hierarchical clustering

3.3 A database of reported trades on CDS indices

Besides the information extracted in the messages received from CDS market-makers, a good market monitoring should also include a real-time feed from the Swap Execution Facilities (SEFs) and the trades that other market participants have done and reported. In this section, we present a dataset of reported trades on CDS indices. This dataset has been created by financial regulations following the financial crisis of 2007-2008 where the opacity of derivatives and swaps markets were largely incriminated. Though these regulations have incurred large operational costs for financial institutions and their clients to report correctly all information required, we think that these data are under-exploited (perhaps due to a difficult and poorly documented access). For now, a curated version of the data is freely and readily available at `www.otcstreaming.com`. We have started to investigate this dataset, and presented our first findings at the XVIII workshop on quantitative finance `https://sites.google.com/site/qfw2017/`. Though this dataset can be valuable on a standalone basis, it could be even more interesting when coupled to other datasets such as CDS quotes. It remains to be determined whether it is really the case.

3.3.1 Regulatory context of swap data repositories

First sentence of the G20 Leaders Statement of the Pittsburgh Summit (September 24-25, 2009) preamble:

> We meet in the midst of a critical transition from crisis to recovery to turn the page on an era of irresponsibility and to adopt a set of policies, regulations and reforms to meet the needs of the 21st century global economy.

More precisely, they propose

> strengthening prudential oversight, improving risk management, strengthening transparency, promoting market integrity, establishing supervisory colleges, and reinforcing international cooperation.

They have

> enhanced and expanded the scope of regulation and oversight, with tougher regulation of over-the-counter (OTC) derivatives.

G20 leaders have decided that:

> All standardized OTC derivative contracts should be traded on exchanges or electronic trading platforms, where appropriate, and cleared through central counterparties by end-2012 at the latest. OTC derivative contracts should be reported to trade repositories. Non-centrally cleared contracts should be subject to higher capital requirements. We ask the Financial Stability Board (FSB) and its relevant members to assess regularly implementation and whether it is sufficient to improve transparency in the derivatives markets, mitigate systemic risk, and protect against market abuse.

These decisions were implemented by the Dodd-Frank Wall Street Reform and Consumer Protection Act (Dodd-Frank Act), July 21, 2010, in the United States, and by the European Market Infrastructure Regulation (EMIR), July 4, 2012, in Europe. EMIR was a major development which enabled the European Union to deliver the G20 commitments on OTC derivatives agreed in Pittsburgh in September 2009, nearly two years after the US.

Swap data repositories (SDRs) are entities created by the Dodd-Frank Act in order to provide a central facility for swap data reporting and recordkeeping. Under the Dodd-Frank Act, all swaps, whether cleared or uncleared, are required to be reported to registered SDRs. SDRs are required to register with the Commodity Futures Trading Commission (CFTC) and comply with rules promulgated by the CFTC, including real-time public reporting of swap transaction and pricing data `www.cftc.gov/industryoversight/datarepositories/index.htm`. These electronic platforms, acting as authoritative registries of key information regarding open OTC derivatives trades, are thought to provide an effective tool for mitigating the inherent opacity of OTC derivatives markets. Several firms are currently registered as SDRs in the US and in Europe.

Concerning credit default swaps, DTCC (Depository Trust & Clearing Corp.) and BSDR LLC (Bloomberg) are the major SDRs, and in practice get all the trades. For now only CDS indices are reported. However, starting the 3rd January 2018, CDS single-names will also be reported in the "Approved Publication Arrangement" (European equivalent of the SDRs) according to MiFIR regulation.

In the next section, we briefly analyze the data that have been recorded since 2012-01-03 on DTCC, and since 2014-05-30 on Bloomberg SDRs (cf. Figure 3.26 for the 5-year iTraxx Europe Main Index ITXEB historical data).

3.3.2 Descriptive statistics of the dataset

The dataset comprises:

- more than 600,000 transactions registered (as of Dec. 2016)
- 4 sources of data: 'Bloomberg SEF', 'Bloomberg OTC', 'Dtcc SEF', 'Dtcc OTC'
- 14 different CDS indices (some of them are not traded/liquid anymore)

Numerical experiments (Python Notebooks) and SDRs historical data access through APIs are made available for reproducible research, and further investigation (`www.otcstreaming.com`).

Statistics of traded volumes

We give here some simple statistics about the volumes traded on the main credit indices. All the graphs can be reproduced using code and data available at `http://public.otcstreaming.com/tech/articles`.

List of CDS indices (Markit defines and owns the indices), and their definition:

- ITXEB, the iTraxx Europe index comprises 125 equally-weighted European names
- ITXEX, the iTraxx Crossover comprises the 75 most liquid sub-investment grade entities
- ITXES, the iTraxx Europe Senior Financials index comprises 25 financial entities from the iTraxx Europe index referencing senior debt
- ITXEU, the iTraxx Europe Subordinated Financials index comprises 25 financial entities from the iTraxx Europe index referencing subordinated debt
- ITXEE, the iTraxx Corp CEEMEA index comprises 25 of the most liquid corporate and quasi-sovereign entities from Central & Eastern European, Middle Eastern and African countries
- ITXHV, the iTraxx Main HiVol index comprises 30 entities with the widest 5-year CDS spreads from the iTraxx Europe Non-Financials index
- ITXSW, the iTraxx Western Europe Sovereign index comprises the 15 most actively traded contracts from Western Europe sovereign credit default swaps
- ITXAA, the iTraxx Australia index comprises 25 of the most liquid Australian entities with investment grade credit ratings

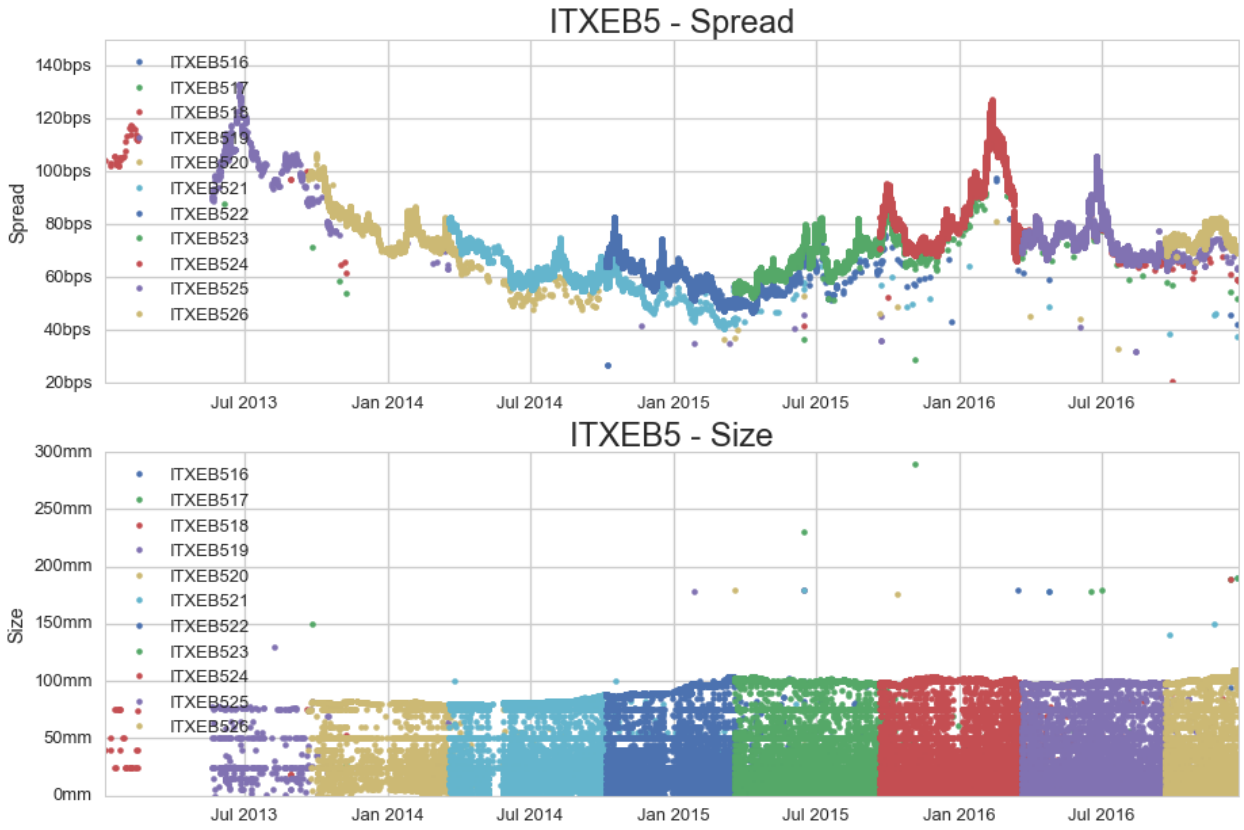


Figure 3.26: Each reported trade is represented by a colored dot. The color indicates the index vintage (which is updated every 6 months) membership.

- ITXAJ, the iTraxx Japan index comprises 50 of the most liquid Japanese entities with investment grade credit ratings
- ITXAG, the iTraxx Asia Ex-Japan index comprises 40 of the most liquid Asian entities with investment grade credit ratings
- CDXIG, the CDX IG index comprises 125 of the most liquid North American entities with investment grade credit ratings
- CDXHY, the CDS HY index comprises 100 liquid North American entities with high yield credit ratings
- CDXEM, the CDX EM index comprises sovereign issuers from Latin America, Eastern Europe, the Middle East, Africa and Asia
- CDXHV, the CDX HiVol index comprises 30 entities in the CDX IG index with the widest 5-year average CDS spreads over the last 90 days prior to the CDX HiVol index composition

Not all of these indices are liquid. Some of them (ITXEE, ITXHV, ITXSW, CDXHV) are not traded any longer. In Table 3.1, we show how many trades occur during a trading day, in average.

Table 3.1: Average number of trades and traded volume for a day

Market	Index	number of trades	traded volume
Europe	ITXEB	121	4687 mm EUR
	ITXEX	122	1717 mm EUR
	ITXES	42	1188 mm EUR
	ITXEU	10	158 mm EUR
	ITXEE	$\ll 1$	
	ITXHV	$\ll 1$	
	ITXSW	$\ll 1$	
Asia	ITXAA	6	123 mm USD
	ITXAJ	7	11703 mm JPY
	ITXAG	8	161 mm USD
America	CDXIG	177	9662 mm USD
	CDXHY	189	3458 mm USD
	CDXEM	45	691 mm USD
	CDXHV	$\ll 1$	

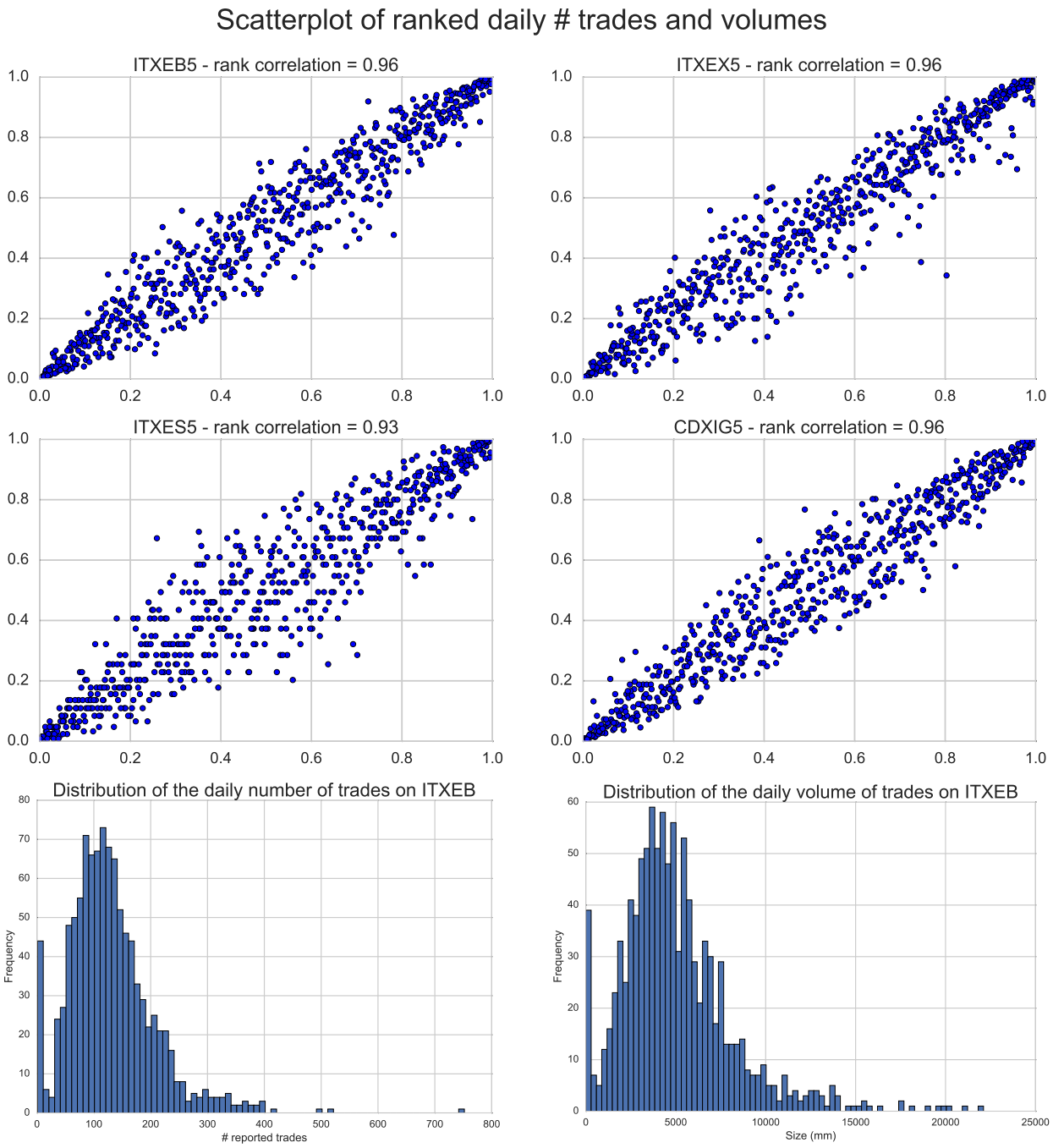
Since the mean is only a rough indicator of the whole distribution. We also display below the full distribution for ITXEB. We notice (cf. Figure 3.27) that though the distributions are skewed, the mean is meaningful and roughly corresponds to the mode of the distribution.

In Table 3.1 and Figure 3.27 data have been aggregated by trading day. In Figure 3.28, we look for seasonality during the trading day (by aggregating data by hours), and throughout the week (by aggregating data by weekdays).

We show in Figure 3.29 the distribution of trades by trade-size. For each index, most of the trades concentrate on a few standard sizes (10, 25, 50, 100mm).

A marked point process

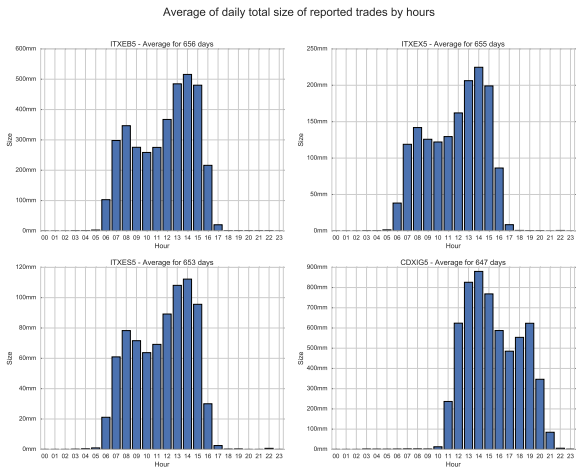
Reported trades arrive in irregular time intervals, while standard econometric techniques are based on fixed time interval analysis. As we have done previously in our simple statistical



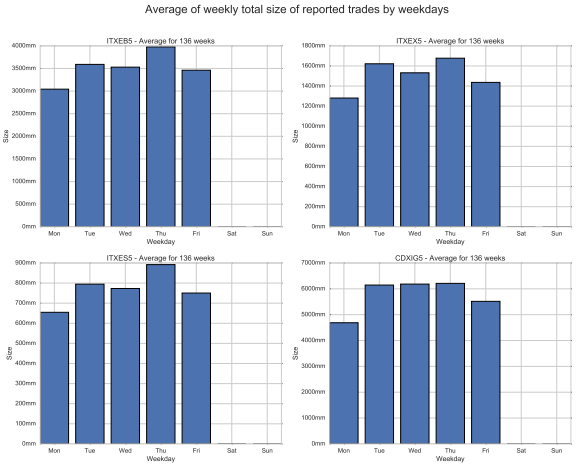
Distribution of the daily number of trades on ITXEB

Distribution of the daily volume of trades on ITXEB

Figure 3.27: (top) Correlation between volumes and number of trades: The daily number of trades and the total volume reported each day are very correlated (about 0.95 for all indices); (below) the distributions are skewed: the volume distribution is more skewed than the number of trades one; (left) number of trades; (right) total volume traded throughout the day



Choosing 'London time' we can see that the most active trading hours are between 2pm and 3pm when both US and European markets are open. That is, in the afternoon for European indices (ITXEB, ITXEX, ITXES), but in the morning (local time) for the US index CDXIG.



One may think that Casual Friday is the coolest day of the week, and that most of the trading have already been done. This prejudice is just plainly wrong according to the SDRs data, and pictured by the graphs below. If there is such a thing as a low activity day, then it is definitely Monday!

Figure 3.28: (left) Average distribution of the traded volume over the day; (right) Average distribution of the traded volume over the week.

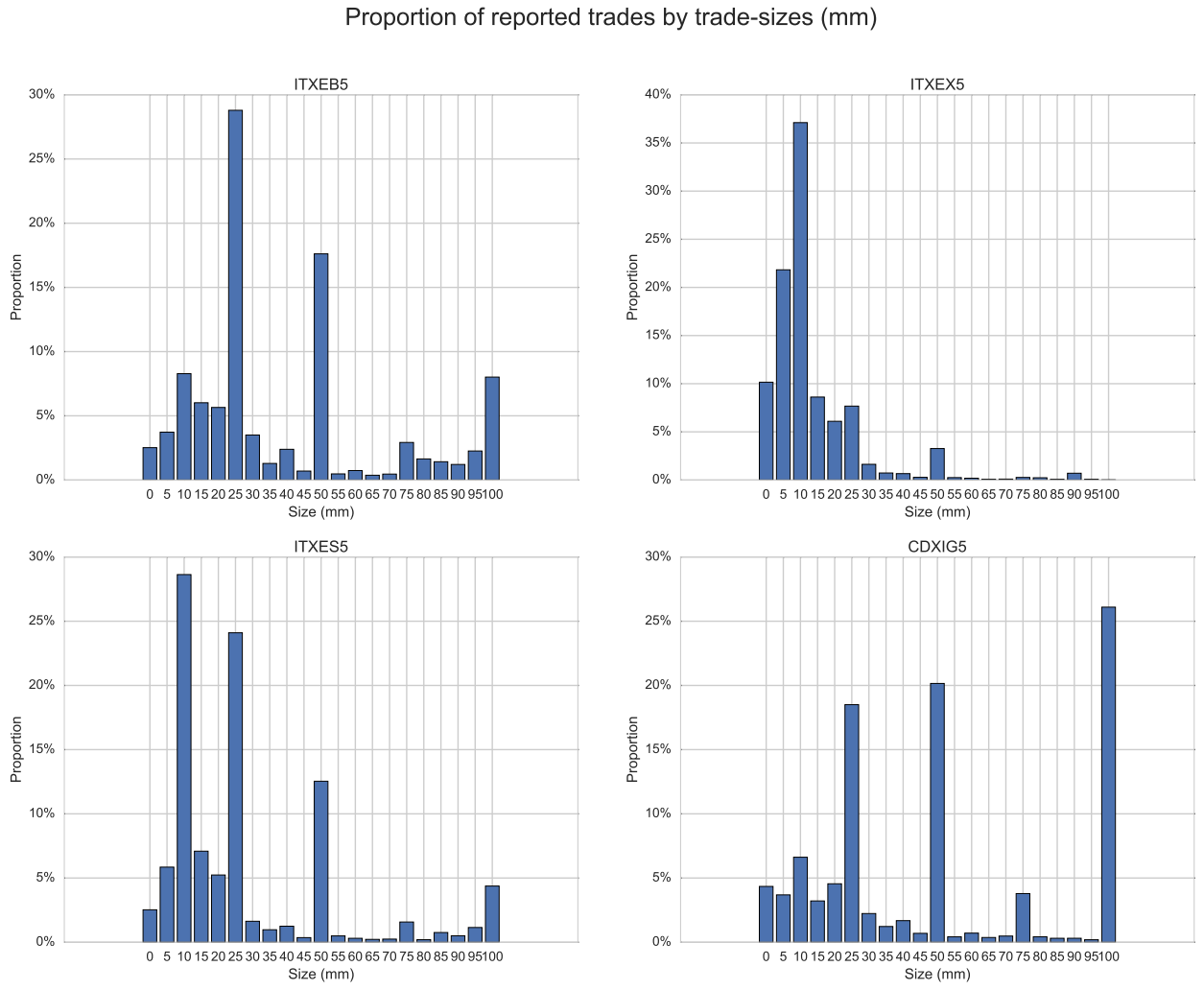


Figure 3.29: We can see that ITXEX usually trades in small sizes (more than 1/3 trades are 10mm) whereas CDXIG trades are usually bigger (more than 1/4 trades are over 100mm).

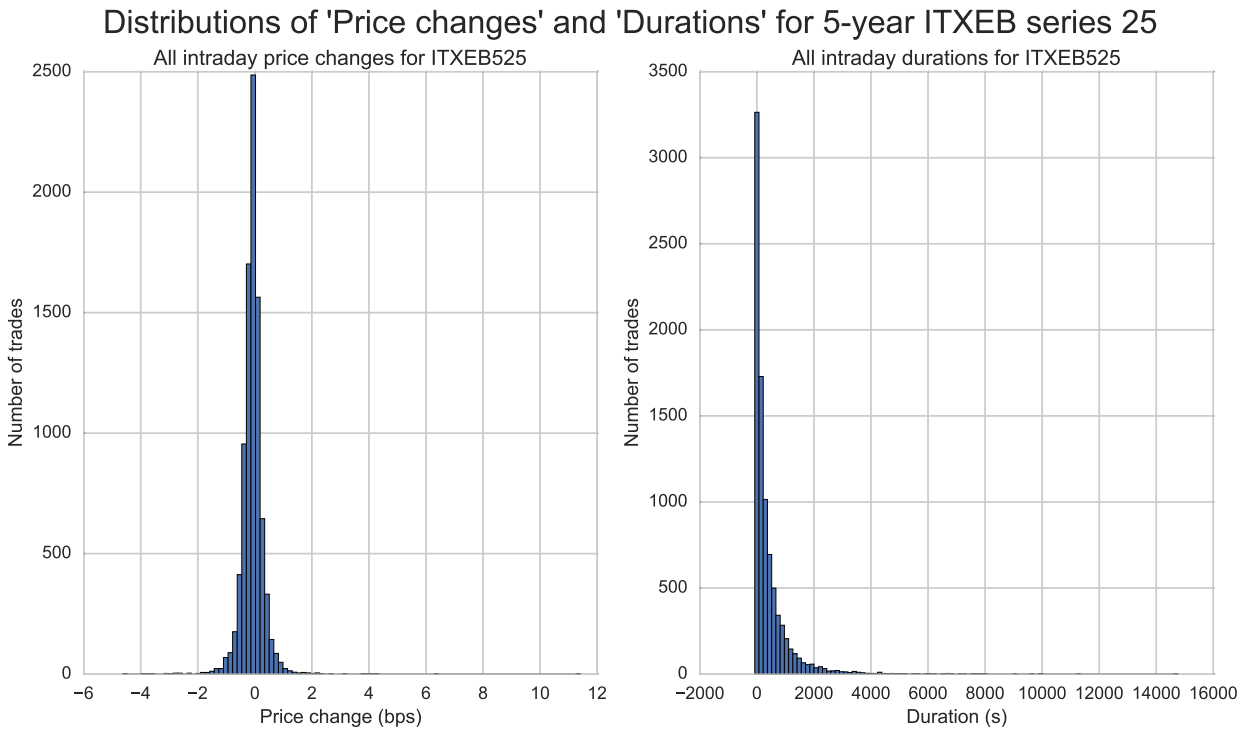


Figure 3.30: Distributions of intraday price changes and durations for ITXEB525

analysis of the traded volumes, we can alleviate this problem by aggregating transaction data. However, the choice of a proper time interval is not obvious: Too short, one can introduce biases due to empty intervals; Too long, much of the information is lost. The 'optimal' length of the interval may also depend on several seasonality factors (as we have seen in Figure 3.28).

For ITXEB525, considering all the intraday data while it has been *on-the-run*, we find that:

- minimum time between two successive trades (duration) is 0 seconds
- maximum duration is 14843 seconds (4 hours 7 minutes 23 seconds)
- median duration is 242 seconds (4 minutes 2 seconds)
- average duration is 487 seconds (8 minutes 7 seconds), with standard deviation of 761 seconds (12 minutes 41 seconds)

Distributions of the intraday price changes and durations for this index (5-year series 25) while being *on-the-run* are displayed in Figure 3.30.

Part II

Novel contributions to the clustering of financial time series

Chapter 4

Consistency of clustering correlated random variables

4.1 Consistency

Researchers have used from 30 days to several years of daily returns as source data for clustering financial time series based on their correlations. This paper sets up a statistical framework to study the validity of such practices. We first show that clustering correlated random variables from their observed values is statistically consistent. Then, we also give a first empirical answer to the much debated question: How long should the time series be? If too short, the clusters found can be spurious; if too long, dynamics can be smoothed out.

Introduction

Clustering can be informally described as the task of grouping objects in subsets (also named clusters) in such a way that objects in the same cluster are more similar to each other than those in different clusters. Since the clustering task is notably hard to formalize [114], designing a clustering algorithm that solves it perfectly in any cases seems farfetched. However, under strong mathematical assumptions made on the data, desirable properties such as statistical consistency, i.e. more data means more accuracy and in the limit a perfect solution, have been shown: Starting from Hartigan's proof of Single Linkage [95] and Pollard's proof of k-means consistency [173] to recent work such as the consistency of spectral clustering [224], or modified k-means [204, 205]. These research papers assume that N data points are independently sampled from an underlying probability distribution in dimension T fixed. Clusters can be seen as regions of high density. They show that in the large sample limit, $N \to \infty$, the clustering sequence constructed by the given algorithm converges to a clustering of the whole underlying space. When we consider the clustering of time series, another asymptotics matter: N fixed and $T \to \infty$. Clusters gather objects that behave similarly through time. To the best of our knowledge, much fewer researchers have dealt with this asymptotics: [23] show the consistency of three hierarchical clustering algorithms when dimension T is growing to correctly gather $N = n + m$ observations from a mixture of two T dimensional Gaussian distributions $\mathcal{N}(\mu_1, \sigma_1^2 I_T)$ and $\mathcal{N}(\mu_2, \sigma_2^2 I_T)$. [182, 109] prove the consistency of k-means for clustering processes according to their *distribution*. In this work, motivated by the clustering of financial time series, we will instead consider the consistency of clustering N random variables from their T observations according to their observed *correlations*.

For financial applications, clustering is usually used as a building block before further processing such as portfolio selection [207]. Before becoming a mainstream methodology among practitioners, one has to provide theoretical guarantees that the approach is sound. In this work, we first show that the clustering methodology is theoretically valid, but when working with finite length time series extra care should be taken: Convergence rates depend

on many factors (underlying correlation structure, separation between clusters, underlying distribution of returns) and implementation choice (correlation coefficient, clustering algorithm). Since financial time series are thought to be approximately stationary for short periods only, a clustering methodology that requires a large sample to recover the underlying clusters is unlikely to be useful in practice and can be misleading. In section 4.2, we illustrate on simulated time series the empirical convergence rates achieved by several clustering approaches.

Notations

- $X_1, \ldots, X_N$ univariate random variables
- X_i^t is the t^{th} observation of variable X_i
- $X_i^{(t)}$ is the t^{th} sorted observation of X_i
- F_X is the cumulative distribution function of X
- $\rho_{ij} = \rho(X_i, X_j)$ correlation between X_i, X_j
- $d_{ij} = d(X_i, X_j)$ distance between X_i, X_j
- $D_{ij} = D(C_i, C_j)$ distance between clusters C_i, C_j
- $P_k = \{C_1^{(k)}, \ldots, C_{l_k}^{(k)}\}$ is a partition of $X_1, \ldots, X_N$
- $C^{(k)}(X_i)$ denotes the cluster of X_i in partition P_k
- $\|\Sigma\|_\infty = \max_{ij} \Sigma_{ij}$
- $X = O_p(k)$ means X/k is stochastically bounded, i.e. $\forall \varepsilon > 0, \exists M > 0, P(|X/k| > M) < \varepsilon$.

4.1.1 The Hierarchical Correlation Block Model

Stylized facts about financial time series

Since the seminal work in [133], it has been verified several times for different markets (e.g. stocks, forex, credit default swaps [137]) that price time series of traded assets have a hierarchical correlation structure. Another well-known stylized fact is the non-Gaussianity of daily asset returns [54]. These empirical properties motivate both the use of alternative correlation coefficients described in section 4.1.1 and the definition of the Hierarchical Correlation Block Model (HCBM) presented in section 4.1.1.

Dependence and correlation coefficients

The most common correlation coefficient is the Pearson correlation coefficient defined by

$$\rho(X, Y) = \frac{\mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y]}{\sqrt{\mathbf{E}[X^2] - \mathbf{E}[X]^2}\sqrt{\mathbf{E}[Y^2] - \mathbf{E}[Y]^2}} \tag{4.1}$$

which can be estimated by

$$\hat{\rho}(X, Y) = \frac{\sum_{t=1}^{T}(X^t - \overline{X})(Y^t - \overline{Y})}{\sqrt{\sum_{t=1}^{T}\left(X^t - \overline{X}\right)^2}\sqrt{\sum_{t=1}^{T}\left(Y^t - \overline{Y}\right)^2}} \tag{4.2}$$

where $\overline{X} = \frac{1}{T}\sum_{t=1}^{T} X^t$ is the empirical mean of X. This coefficient suffers from several drawbacks: it only measures linear relationship between two variables; it is not robust to

noise and may be undefined if the distribution of one of these variables have infinite second moment. More robust correlation coefficients are copula-based dependence measures such as Spearman's rho

$$\begin{aligned} \rho_S(X,Y) &= 12\int_0^1\int_0^1 C(u,v)dudv - 3 && (4.3)\\ &= 12\,\mathbf{E}\left[F_X(X), F_Y(Y)\right] - 3 && (4.4)\\ &= \rho\left(F_X(X), F_Y(Y)\right) && (4.5) \end{aligned}$$

and its statistical estimate

$$\hat{\rho}_S(X,Y) = 1 - \frac{6}{T(T^2-1)}\sum_{t=1}^{T}\left(X^{(t)} - Y^{(t)}\right)^2. \quad (4.6)$$

These correlation coefficients are robust to noise (since rank statistics normalize outliers) and invariant to monotonous transformations of the random variables (since copula-based measures benefit from the probability integral transform $F_X(X) \sim \mathcal{U}[0,1]$).

The HCBM model

We assume that the N univariate random variables $X_1, \ldots, X_N$ follow a Hierarchical Correlation Block Model (HCBM). This model consists in correlation matrices having a hierarchical block structure [9], [117]. Each block corresponds to a correlation cluster that we want to recover with a clustering algorithm. In Fig. 4.1, we display a correlation matrix from the HCBM. Notice that in practice one does not observe the hierarchical block diagonal structure displayed in the left picture, but a correlation matrix similar to the one displayed in the right picture which is identical to the left one up to a permutation of the data. The HCBM defines a set of nested partitions $\mathcal{P} = \{P_0 \supseteq P_1 \supseteq \ldots \supseteq P_h\}$ for some $h \in [1, N]$, where P_0 is the trivial partition, the partitions $P_k = \{C_1^{(k)}, \ldots, C_{l_k}^{(k)}\}$, and $\bigsqcup_{i=1}^{l_k} C_i^{(k)} = \{X_1, \ldots, X_N\}$. For all $1 \leq k \leq h$, we define $\underline{\rho}_k$ and $\overline{\rho}_k$ such that for all $1 \leq i, j \leq N$, we have $\underline{\rho}_k \leq \rho_{ij} \leq \overline{\rho}_k$ when $C^{(k)}(X_i) = C^{(k)}(X_j)$ and $C^{(k+1)}(X_i) \neq C^{(k+1)}(X_j)$, i.e. $\underline{\rho}_k$ and $\overline{\rho}_k$ are the minimum and maximum correlation respectively within all the clusters $C_i^{(k)}$ in the partition P_k at depth k. In order to have a proper nested correlation hierarchy, we must have $\overline{\rho}_k < \underline{\rho}_{k+1}$ for all k. Depending on the context, it can be a Spearman or Pearson correlation matrix.

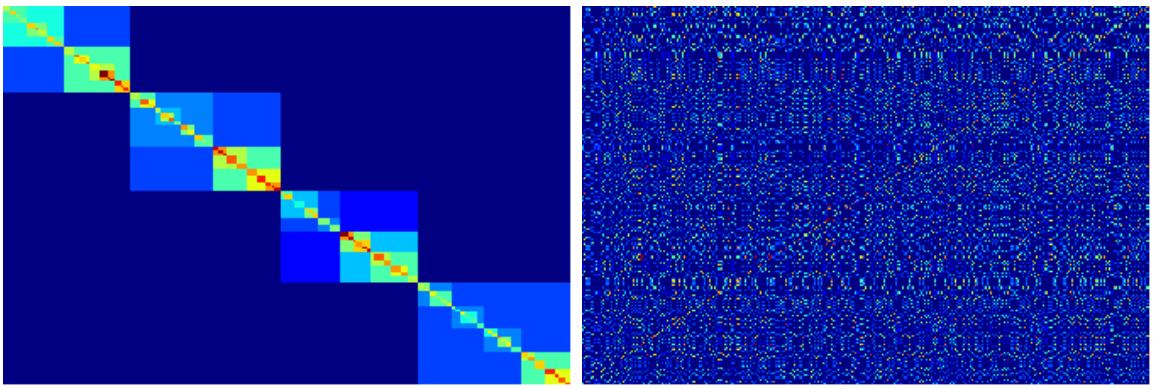


Figure 4.1: (left) hierarchical correlation block model; (right) observed correlation matrix (following the HCBM) identical to the left one up to a permutation of the data

Without loss of generality and for ease of demonstration we will consider the one-level HCBM with K blocks of size $n_1, \ldots, n_K$ such that $\sum_{i=1}^{K} n_i = N$. We explain later how to extend the results to the general HCBM. We also consider the associated distance matrix d, where $d_{ij} = \frac{1-\rho_{ij}}{2}$. In practice, clustering methods are applied on statistical estimates of the distance matrix d, i.e. on $\hat{d}_{ij} = d_{ij} + \epsilon_{ij}$, where ϵ_{ij} are noises resulting from the statistical estimation of correlations.

Table 4.1: Many well-known hierarchical agglomerative clustering algorithms are members of the Lance-Williams family, i.e. the distance between clusters can be written:
$D(C_i \cup C_j, C_k) = \alpha_i D_{ik} + \alpha_j D_{jk} + \beta D_{ij} + \gamma |D_{ik} - D_{jk}|$

	α_i	β	γ
Single	$1/2$	0	-1/2
Complete	$1/2$	0	$1/2$
Average	$\frac{\lvert C_i \rvert}{\lvert C_i \rvert + \lvert C_j \rvert}$	0	0
McQuitty	$1/2$	0	0
Median	$1/2$	-1/4	0
Centroid	$\frac{\lvert C_i \rvert}{\lvert C_i \rvert + \lvert C_j \rvert}$	$-\frac{\lvert C_i \rvert \lvert C_j \rvert}{(\lvert C_i \rvert + \lvert C_j \rvert)^2}$	0
Ward	$\frac{\lvert C_i \rvert + \lvert C_k \rvert}{\lvert C_i \rvert + \lvert C_j \rvert + \lvert C_k \rvert}$	$-\frac{\lvert C_k \rvert}{\lvert C_i \rvert + \lvert C_j \rvert + \lvert C_k \rvert}$	0

Clustering methods

Algorithms of interest

Many paradigms exist in the literature for clustering data. We consider in this work only hard (in opposition to soft) clustering methods, i.e. algorithms producing partitions of the data (in opposition to methods assigning several clusters to a given data point). Within the hard clustering family, we can classify for instance these algorithms in hierarchical clustering methods (yielding nested partitions of the data) and flat clustering methods (yielding a single partition) such as k-means.

We will consider the infinite Lance-Williams family which further subdivides the hierarchical clustering since many of the popular algorithms such as Single Linkage, Complete Linkage, Average Linkage (UPGMA), McQuitty's Linkage (WPGMA), Median Linkage (WPGMC), Centroid Linkage (UPGMC), and Ward's method are members of this family (cf. Table 4.1 [153]). It will allow us a more concise and unified treatment of the consistency proofs for these algorithms. Interesting and recently designed hierarchical agglomerative clustering algorithms such as Hausdorff Linkage [12] and Minimax Linkage [4] do not belong to this family [17], but their linkage functions share a convenient property for cluster separability.

Separability conditions for clustering

In our context the distances between the points we want to cluster are random and defined by the estimated correlations. However by definition of the HCBM, each point X_i belongs to exactly one cluster $C^{(k)}(X_i)$ at a given depth k, and we want to know under which condition on the distance matrix we will find the correct clusters defined by P_k. We call these conditions the separability conditions. A separability condition for the points $X_1, \ldots, X_N$ is a condition on the distance matrix of these points such that if we apply a clustering procedure whose input is the distance matrix, then the algorithm yields the correct clustering $P_k = \{C_1^{(k)}, \ldots, C_{l_k}^{(k)}\}$, . For example, for $\{X_1, X_2, X_3\}$ if we have $C(X_1) = C(X_2) \neq C(X_3)$ in the one-level two-block HCBM, then a separability condition is $d_{1,2} < d_{1,3}$ and $d_{1,2} < d_{2,3}$.

Separability conditions are deterministic and depend on the algorithm used for clustering. They are generic in the sense that for any sets of points that satisfy the condition the algorithm will separate them in the correct clusters. In the Lance-Williams algorithm framework [51], they are closely related to "space conserving" properties of the algorithm and in particular on the way the distances between clusters change during the clustering process.

Space-conserving algorithms

In [51], the authors define what they call a semi-space-conserving algorithm.

Definition 1 (Semi-space-conserving algorithms)**.** *An algorithm is semi-space-conserving if for all clusters C_i, C_j, and C_k,*

$$D(C_i \cup C_j, C_k) \in [\min(D_{ik}, D_{jk}), \max(D_{ik}, D_{jk})]$$

Among the Lance-Williams algorithms we study here, Single, Complete, Average and McQuitty algorithms are semi-space-conserving. Although Chen and Van Ness only considered Lance-Williams algorithms the definition of a space conserving algorithm is useful for any agglomerative hierarchical algorithm. An alternative formulation of the semi-space-conserving property is:

Definition 2 (Space-conserving algorithms)**.** *A linkage agglomerative hierarchical algorithm is space-conserving if* $D_{ij} \in \left[\min_{x \in C_i, y \in C_j} d(x,y), \max_{x \in C_i, y \in C_j} d(x,y)\right]$.

Such an algorithm does not "distort" the space when points are clustered which makes the sufficient separability condition easier to get. For these algorithms the separability condition does not depend on the size of the clusters.

The following two propositions are easy to verify.

Property 1. *The semi-space-conserving Lance-Williams algorithms are space-conserving.*

Property 2. *Minimax linkage and Hausdorff linkage are space-conserving.*

For space-conserving algorithms we can now state a sufficient separability condition on the distance matrix.

Property 3. *The following condition is a separability condition for space-conserving algorithms:*

$$\max_{\substack{1 \le i,j \le N \\ C(i)=C(j)}} d(X_i, X_j) \; < \min_{\substack{1 \le i,j \le N \\ C(i) \neq C(j)}} d(X_i, X_j) \tag{S1}$$

The maximum distance is taken over any two points in a same cluster (intra) and the minimum over any two points in different clusters (inter).

Proof. Consider the set $\{d_{ij}^s\}$ of distances between clusters after s steps of the clustering algorithm (therefore $\{d_{ij}^0\}$ is the initial set of distances between the points). Denote $\{d_{inter}^s\}$ (resp. $\{d_{intra}^s\}$) the sets of distances between subclusters belonging to different clusters (resp. the same cluster) at step s. If the separability condition is satisfied then we have the following inequalities:

$$\min d_{intra}^0 \le \max d_{intra}^0 < \min d_{inter}^0 \le \max d_{inter}^0 \tag{S2}$$

Then the separability condition implies that the separability condition S2 is verified for all step s because after each step the updated intra distances are in the convex hull of the intra distances of the previous step and the same is true for the inter distances. Moreover since S2 is verified after each step, the algorithm never links points from different clusters and the proposition entails. □

Ward algorithm

The Ward algorithm is a space-dilating Lance-Williams algorithm: $D(C_i \cup C_j, C_k) > \max(D_{ik}, D_{jk})$. This is a more complicated situation because the structure

$$\min d_{inter} < \max d_{inter} < \min d_{intra} < \max d_{intra}$$

is not necessarily preserved under the condition $\max d_{inter}^0 < \min d_{intra}^0$. Points which are not clustered move away from the clustered points. Outliers, which will only be clustered at the very end, will end up close to each other and far from the clustered points. This can lead to wrong clusters. Therefore a generic separability condition for Ward needs to be stronger and account for the distortion of the space. Since the distortion depends on the number of steps the algorithm needs, the separability condition depends on the size of the clusters.

Property 4 (Separability condition for Ward)**.** *The separability condition for Ward reads:*

$$n[\max d^0_{intra} - \min d^0_{intra}] < [\min d^0_{inter} - \min d^0_{intra}]$$

where $n = \max_i n_i$ *is the size of the largest cluster.*

Proof. Let A and B be two subsets of the N points of size a and b respectively. Then

$$D(A,B) = \frac{ab}{a+b}\left(\frac{2}{ab}\sum_{\substack{i\in A\\ j\in B}} d_{ij} - \frac{1}{a^2}\sum_{\substack{i\in A\\ i'\in A}} d_{ii'} - \frac{1}{b^2}\sum_{\substack{j\in B\\ j'\in B}} d_{jj'}\right)$$

is a linkage function for the Ward algorithm. To ensure that the Ward algorithm will never merge the wrong subsets it is sufficient that for any sets A and B in a same cluster, and A', B' in different clusters, we have:

$$D(A,B) < D(A',B').$$

Since

$$\begin{cases} D(A,B) \leq n(\max d^0_{intra} - \min d^0_{intra}) + \min d^0_{intra} - 1 \\ D(A',B') \geq (\min d^0_{inter} - \max d^0_{intra}) + \max d^0_{intra} - 1 \end{cases}$$

we obtain the condition:

$$n(\max d^0_{intra} - \min d^0_{intra}) < \min d^0_{inter} - \min d^0_{intra}.$$

□

k-means

The k-means algorithm is not a linkage algorithm. For the k-means algorithm we need a separability condition that ensures that the initialization will be good enough for the algorithm to find the partition. In [182] (Theorem 1), the author proves the consistency of the one-step farthest-point initialization k-means [104] with a distributional distance for clustering processes. The separability condition S1 of Proposition 3 is sufficient for k-means.

4.1.2 Consistency of well-known clustering algorithms

In the previous section we have determined configurations of points such that the clustering algorithm will find the right partition. The proof of the consistency now relies on showing that these configurations are likely. In fact the probability that our points fall in these configurations goes to 1 as $T \to \infty$.

The precise definition of what we mean by consistency of an algorithm is the following:

Definition 3 (Consistency of a clustering algorithm)**.** *Let* $(X_1^t, \ldots, X_N^t)$, $t = 1, \ldots, T$, *be* N *univariate random variables observed* T *times. A clustering algorithm* $\mathcal{A}$ *is consistent with respect to the Hierarchical Correlation Block Model (HCBM) defining a set of nested partitions* $\mathcal{P}$ *if the probability that the algorithm* $\mathcal{A}$ *recovers all the partitions in* $\mathcal{P}$ *converges to* 1 *when* $T \to \infty$.

As we have seen in the previous section the correct clustering can be ensured if the estimated correlation matrix verifies some separability condition. This condition can be guaranteed by requiring the error on each entry of the matrix $\hat{R}_T$ to be smaller than the contrast, i.e. $\frac{\underline{\rho}_1 - \overline{\rho}_0}{2}$, on the theoretical matrix R. There are classical results on the concentration properties of estimated correlation matrices such as:

Theorem 5 (Concentration properties of the estimated correlation matrices [131]). *If Σ and $\hat{\Sigma}$ are the population and empirical Spearman correlation matrix respectively, then with probability at least $1 - \frac{1}{T^2}$, for $N \geq \frac{24}{\log T} + 2$, we have*

$$\|\hat{\Sigma} - \Sigma\|_\infty \leq 24\sqrt{\frac{\log N}{T}}$$

The concentration bounds entails that if $T \gg \log(N)$ then the clustering will find the correct partition because the clusters will be sufficiently separated with high probability. In financial applications of clustering, we need the error on the estimated correlation matrix to be small enough for relatively short time-windows. However there is a dimensional dependency of these bounds [210] that make them uninformative for realistic values of N and T in financial applications, but there is hope to improve the bounds using the special structure of HCBM correlation matrices.

From the one-level to the general HCBM

To go from the one-level HCBM to the general case we need to get a separability condition on the nested partition model. For both space-conserving algorithms and the Ward algorithm, this is done by requiring the corresponding separability condition for each level of the hierarchy.

For all $1 \leq k \leq h$, we define $\underline{d}_k$ and $\overline{d}_k$ such that for all $1 \leq i, j \leq N$, we have $\underline{d}_k \leq d_{ij} \leq \overline{d}_k$ when $C^{(k)}(X_i) = C^{(k)}(X_j)$ and $C^{(k+1)}(X_i) \neq C^{(k+1)}(X_j)$. Notice that $\underline{d}_k = (1 - \overline{\rho}_k)/2$ and $\overline{d}_k = (1 - \underline{\rho}_k)/2$.

Property 6. *[Separability condition for space-conserving algorithms in the case of nested partitions] The separability condition reads:*

$$\overline{d}_h < \underline{d}_{h-1} < \ldots < \overline{d}_{k+1} < \underline{d}_k < \ldots < \underline{d}_1.$$

This condition can be guaranteed by requiring the error on each entry of the matrix $\hat{\Sigma}$ to be smaller than the lowest contrast. Therefore the maximum error we can have for space-conserving algorithms on the correlation matrix is

$$\|\Sigma - \hat{\Sigma}\|_\infty < \min_k \frac{\underline{\rho}_{k+1} - \overline{\rho}_k}{2}.$$

Property 7. *[Separability condition for the Ward algorithm in the case of nested partitions] Let n_k be the size of the largest cluster at the level k of the hierarchy.*

The separability condition reads:

$$\forall k \in \{1, \ldots, h\}, \qquad n_k(\overline{d}_k - \underline{d}_h) < \underline{d}_{k-1} - \underline{d}_h$$

Therefore the maximum error we can have for space-conserving algorithms on the correlation matrix is

$$\|\Sigma - \hat{\Sigma}\|_\infty < \min_k \frac{\overline{\rho}_h - \overline{\rho}_{k-1} - n_k(\overline{\rho}_h - \underline{\rho}_k)}{1 + 2n_k},$$

where n_k is the size of the largest cluster at the level k of the hierarchy.

We finally obtain consistency for the presented algorithms with respect to the HCBM from the previous concentration results.

4.2 Empirical rates of convergence

We have shown in the previous sections that clustering correlated random variables is consistent under the hierarchical correlation block model. This model is supported by many

empirical studies [133] where the authors scrutinize time series of returns for several asset classes. However, it was also noticed that the correlation structure is not fixed and tends to evolve through time. This is why, besides being consistent, the convergence of the methodology needs to be fast enough for the underlying clustering to be accurate. For now, theoretical bounds such as the ones obtained in Theorem 5 are uninformative for realistic values of N and T. For example, for $N = 265$ and $T = 2500$ (roughly 10 years of historical daily returns) with a separation between clusters of $d = 0.2$, we are confident with probability greater than $1 - 2N^2e^{-Td^2/24} \approx -2176$ that the clustering algorithm has recovered the correct clusters. These bounds will eventually converge to 0 with rate $O_P(\sqrt{\log N}/\sqrt{T})$. In addition, the convergence rates also depend on many factors, e.g. the number of clusters, their relative sizes, their separations, whose influence is very specific to a given clustering algorithm, and thus difficult to consider in a theoretical analysis.

To get an idea of the minimal amount of data one should use in applications to be confident with the clustering results, we suggest to design realistic simulations of financial time series and determine the sample critical size from which the clustering approach "always" recovers the underlying model. We illustrate such an empirical study in the following section.

Financial time series models

For the simulations, implementation and tutorial available at `www.datagrapple.com/Tech`, we will consider two models:

- The standard but debated model of quantitative finance, the Gaussian random walk model whose increments are realizations from a N-variate Gaussian: $X \sim \mathcal{N}(0, \Sigma)$.

The Gaussian model does not generate heavy-tailed behavior (strong unexpected variations in the price of an asset) which can be found in many asset returns [54] nor does it generate tail-dependence (strong variations tend to occur at the same time for several assets). This oversimplified model provides an empirical convergence rate for clustering that is unlikely to be exceeded on real data.

- The increments are realizations from a N-variate Student's t-distribution, with degree of freedom $\nu = 3$: $X \sim t_\nu(0, \frac{\nu-2}{\nu}\Sigma)$.

The N-variate Student's t-distribution ($\nu = 3$) captures both the heavy-tailed behavior (since marginals are univariate Student's t-distribution with the same parameter $\nu = 3$) and the tail-dependence. It has been shown that this distribution yields a much better fit to real returns than the Gaussian distribution [97].

The Gaussian and t-distribution are parameterized by a covariance matrix Σ. We define Σ such that the underlying correlation matrix has the structure depicted in Figure 5.7. This correlation structure is inspired by the real correlations between credit default swap assets in the European "investment grade", European "high-yield" and Japanese markets. More precisely, this correlation matrix allows us to simulate the returns time series for $N = 265$ assets divided into

- a "European investment grade" cluster composed of 115 assets, subdivided into
  - 7 industry-specific clusters of sizes 10, 20, 20, 5, 30, 15, 15; the pairwise correlation inside these 7 clusters is 0.7;
- a "European high-yield" cluster composed of 100 assets, subdivided into
  - 7 industry-specific clusters of sizes 10, 20, 25, 15, 5, 10, 15; the pairwise correlation inside these 7 clusters is 0.7;
- a "Japanese" cluster composed of 50 assets whose pairwise correlation is 0.6.

We can then sample time series from these two models.

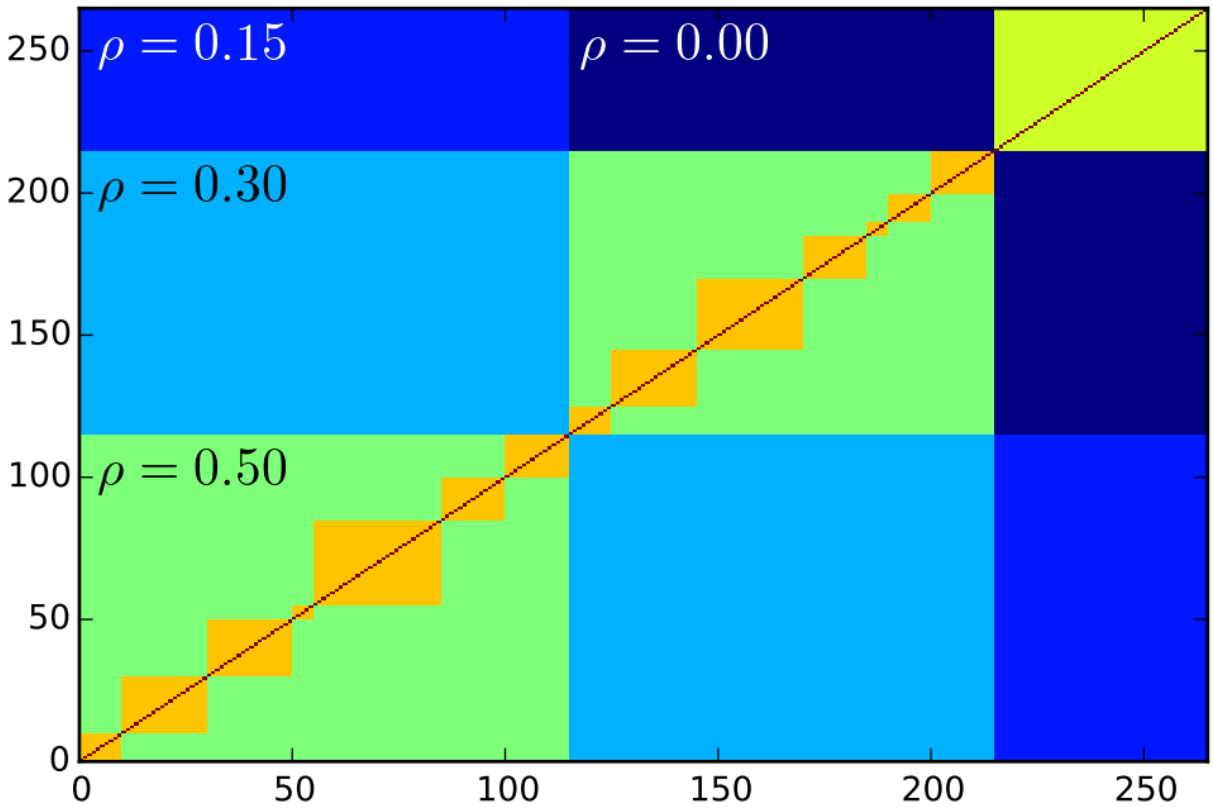


Figure 4.2: Illustration of the correlation structure used for simulations: European assets (numbered $0, \ldots, 214$) are subdivided into 2 clusters which are themselves subdivided into 7 clusters each; Japanese assets (numbered $215 \ldots, 264$) are weakly correlated to the European markets: $\rho = 0.15$ with "investment grade" assets, $\rho = 0.00$ with "high-yield" assets

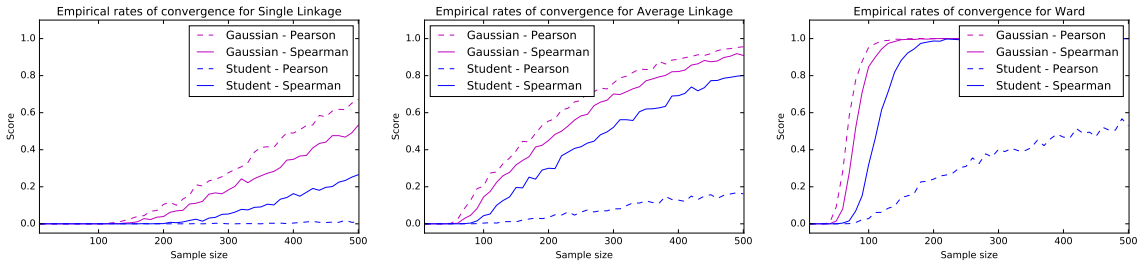


Figure 4.3: Single Linkage (left), Average Linkage (mid), Ward method (right) are used for clustering the simulated time series; Dashed lines represent the ratio of correct clustering over the number of trials (the score displayed in y-axis) when using Pearson coefficient, solid lines for the Spearman one; Magenta lines are used when the underlying model is Gaussian, blue lines for the t-distribution

Experiment: Recovering the initial clusters

For each model, for every T ranging from 10 to 500, we sample $L = 10^3$ datasets of $N = 265$ time series with length T from the model. We count how many times the clustering methodology (here, the choice of an algorithm and a correlation coefficient) is able to recover the underlying clusters defined by the correlation matrix. In Figure 4.3, we display the results obtained using *Single Linkage* (motivated in Mantegna *et al.*'s research [134] by the ultrametric space hypothesis and the related subdominant ultrametric given by the minimum spanning tree), *Average Linkage* (which is used to palliate against the unbalanced effect of Single Linkage, yet unlike Single Linkage, it is sensitive to monotone transformations of the distances d_{ij}) and the *Ward method* leveraging either the Pearson correlation coefficient or the Spearman one.

Conclusions from the empirical study

As expected, the Pearson coefficient yields the best results when the underlying distribution is Gaussian and the worst when the underlying distribution is heavy-tailed. For such elliptical distributions, rank-based correlation estimators are more relevant [130, 93]. Concerning clustering algorithm convergence rates, we find that Average Linkage outperforms Single Linkage for $T \ll N$ and $T \simeq N$. One can also notice that both Single Linkage and Average Linkage have not yet converged after 500 realizations (roughly 2 years of daily returns) whereas the Ward method, which is not mainstream in the econophysics literature, has converged after 250 realizations (about a year of daily returns). Its variance is also much smaller. Based on this empirical study, a practitioner working with $N = 265$ assets whose underlying correlation matrix may be similar to the one depicted in Figure 5.7 should use the Ward + Spearman methodology on a sliding window of length $T = 250$.

Application to credit default swaps

From Hellebore Capital database, we use $N = 561$ daily time series with full history starting from the 01/01/2006, that is $T = 2805$ observations for each entity as of April 2017. We then

- compute an empirical correlation matrix of this sample (displayed in Figure 4.4) that we serialize for visualizing its hierarchical structure (displayed in Figure 4.5). We serialize the correlation matrix by traversing recursively and top down the dendrogram obtained from an agglomerative hierarchical clustering algorithm which allows us to find a good permutation of the rows (we apply the same permutation for the columns), i.e. a kind of 'quicksort' of the rows (and columns);

- compute a 'reasonable' number of clusters (according to a model selection method, e.g. using bootstrapping and a stability score). In this case, we choose $K = 30$ clusters. In Figure 4.6, we overlay the boundaries of these $K = 30$ clusters on the empirical correlation matrix of the whole sample.

- average the correlations inside each of the $K = 30$ clusters (displayed in Figure 4.7).

- build a filtered correlation matrix (displayed in Figure 4.8) which can be used to parameterize a multivariate distribution such as a Gaussian or Student one.

- sample from the previous model, i.e. from a distribution parameterized by the filtered correlation matrix. In Figure 4.9 and Figure 4.10, we show the empirical correlation matrix estimated from a sample of size $T = 250$ (about 1 year of historical daily 'returns') and $T = 2000$ (about 8 years of historical daily 'returns') respectively.

- estimate the $K = 30$ clusters on different samples of increasing sizes. Compare them to the $K = 30$ original clusters encoded in the correlation matrix model using a similarity score. We display the scores obtained in Figure 4.11. We can notice that we can perfectly recover the $K = 30$ original clusters with at least $T = 1000$ observations (about 4 years of historical daily 'returns').

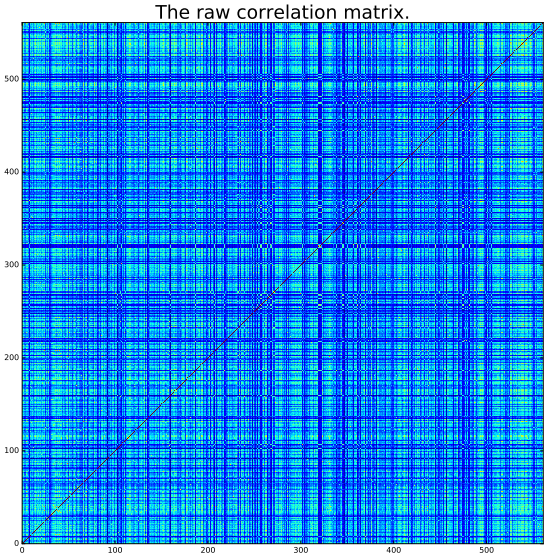


Figure 4.4: Empirical correlations

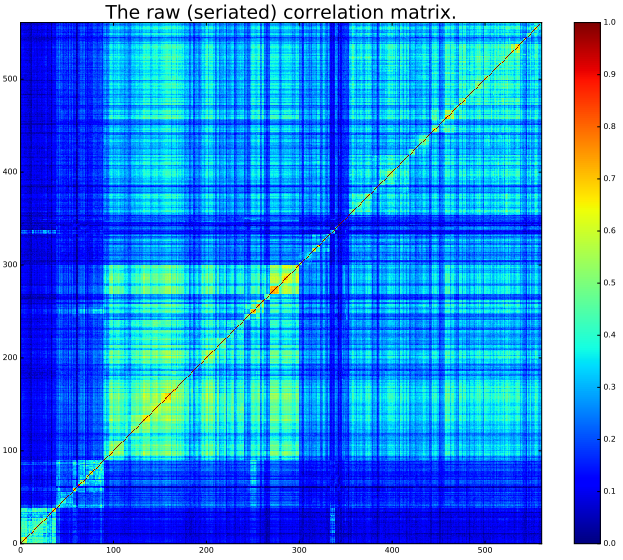


Figure 4.5: Seriated correlations

Discussion

In this contribution, we only show consistency with respect to a model motivated by empirical evidence. *All models are wrong* and this one is no exception to the rule: random walk hypothesis, real correlation matrices are not that "blocky". We identified several theoretical directions for the future:

- The theoretical concentration bounds are not sharp enough for usual values of N, T. Since the intrinsic dimension of the correlation matrices in the HCBM is low, there might be some possible improvements [210].

- "Space-conserving", "space-dilating" is a coarse classification that does not allow to distinguish between several algorithms with different behaviors. Though Single Linkage (which is nearly "space-contracting") and Average Linkage have different convergence rates as shown by the empirical study, they share the same theoretical bounds.

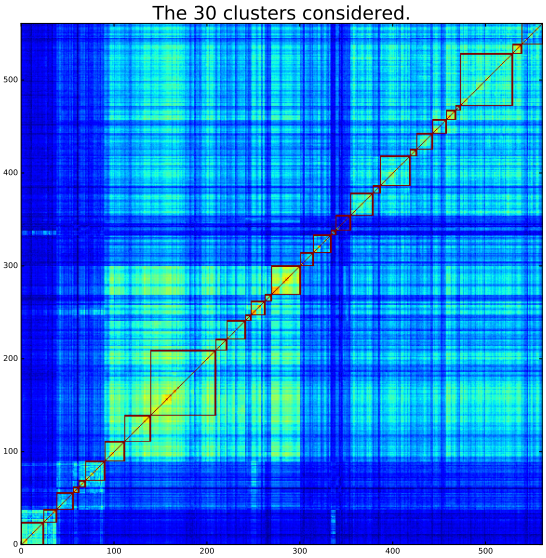


Figure 4.6: Overlaid clusters boundaries

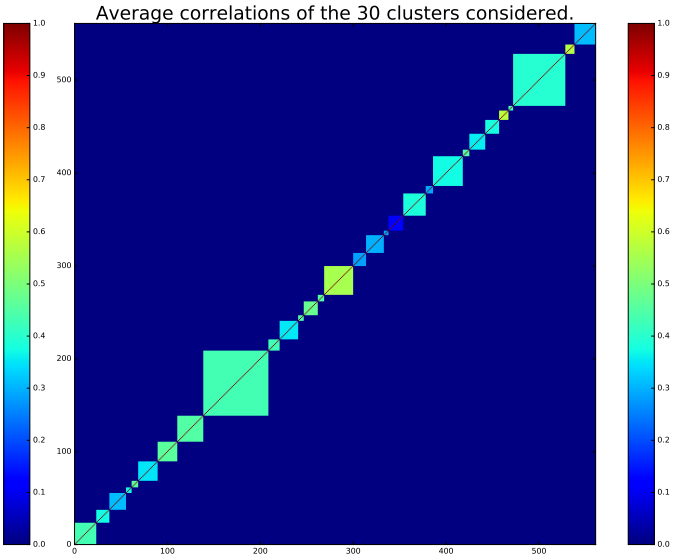


Figure 4.7: Clusters average correlations

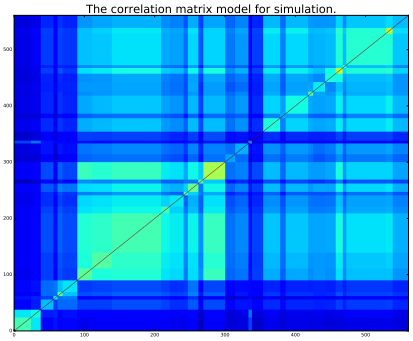


Figure 4.8: Correlations model

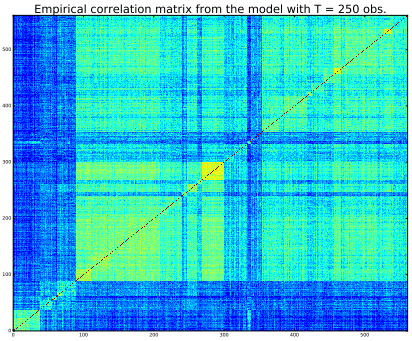


Figure 4.9: $T = 250$

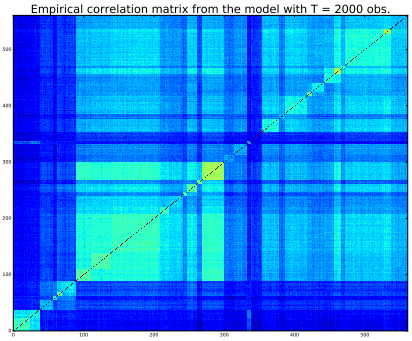


Figure 4.10: $T = 2000$

And also directions for experimental studies:

- It would be interesting to study spectral clustering techniques which are less greedy than the hierarchical clustering algorithms. In [217], authors show that they are less stable with respect to statistical uncertainty than hierarchical clustering. Less stability may imply a slower convergence rate.

- We notice that there are isoquants of clustering accuracy for many sets of parameters, e.g. (N, T), (ρ, T). Such isoquants are displayed in Figure 4.12. Further work may aim at characterizing these curves. We can also observe in Figure 4.12 that for $\rho \leq 0.08$, the critical value for T explodes. It would be interesting to determine this asymptotics as ρ tends to 0.

Finally, we have provided a guideline to help the practitioner set the critical window-size T for a given clustering methodology. One can also investigate which consistent methodology provides the correct clustering the fastest. However, much work remains to understand the convergence behaviors of clustering algorithms on financial time series.

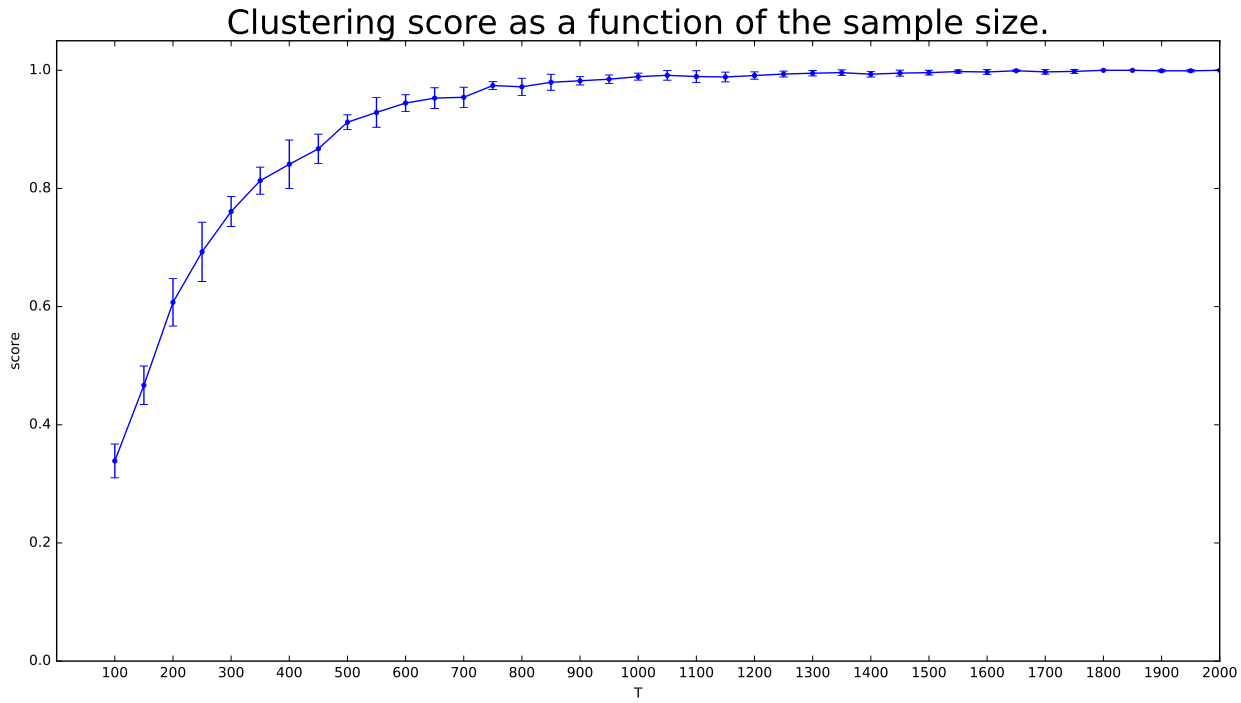


Figure 4.11: The clustering scores indicate that $T = 1000$ is a suitable sample size for being confident in the clustering results.

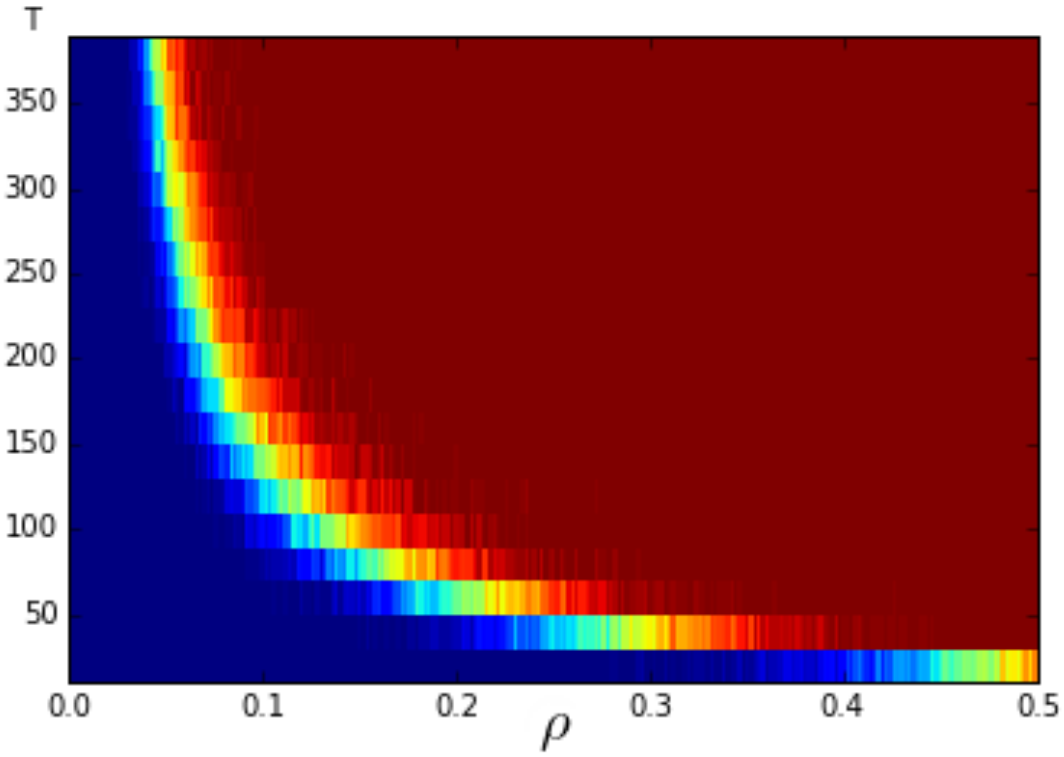


Figure 4.12: Heatmap encoding the ratio of correct clustering over the number of trials (score in $[0, 1]$) for the Ward + Spearman methodology as a function of ρ and T; underlying model is a Gaussian distribution parameterized by a 2-block-uniform-ρ correlation matrix; red color (score = 1) represents a perfect and systematic recovering of the underlying two clusters, deep blue (score = 0) encodes 0 correct clustering; notice the clear-cut isoquants

Chapter 5

Distances between financial time series

5.1 A correlation/distribution distance

Motivation: A better representation which is appropriately taken into account with a suitable distance should imply a more stable and faster convergence clustering.

We introduce a novel non-parametric approach to represent random variables which splits apart dependency and distribution without losing any information. We also propound an associated metric leveraging this representation and its statistical estimate. Besides experiments on synthetic datasets, the benefits of our contribution is illustrated through the example of clustering financial time series, for instance prices from the credit default swaps market. Results are available on the website `www.datagrapple.com` and an IPython Notebook tutorial is available at `www.datagrapple.com/Tech` for reproducible research.

Introduction

Machine learning on time series is a booming field and as such plenty of representations, transformations, normalizations, metrics and other divergences are thrown at disposal to the practitioner. A further consequence of the recent advances in time series mining is that it is difficult to have a sober look at the state of the art since many papers state contradictory claims as described in [66]. To be fair, we should mention that when data, pre-processing steps, distances and algorithms are combined together, they have an intricate behaviour making it difficult to draw unanimous conclusions especially in a fast-paced environment. Restricting the scope of time series to independent and identically distributed (i.i.d.) stochastic processes, we propound a method which, on the contrary to many of its counterparts, is mathematically grounded with respect to the clustering task defined in subsection 5.1. The representation we present in Section 5.1 exploits a property similar to the seminal result of copula theory, namely Sklar's theorem [192]. This approach leverages the specificities of random variables and this way solves several shortcomings of more classical data preprocessing and distances that will be detailed in subsection 5.1. Section 5.1 is dedicated to experiments on synthetic and real datasets to illustrate the benefits of our method which relies on the hypothesis of i.i.d. sampling of the random variables. Synthetic time series are generated by a simple model yielding correlated random variables following different distributions. The presented approach is also applied to financial time series from the credit default swaps market whose prices dynamics are usually modelled by random walks according to the efficient-market hypothesis [80]. This dataset seems more interesting than stocks as credit default swaps are often considered as a gauge of investors' fear, thus time series are subject to more violent moves and may provide more distributional information than the ones from the stock market. We have made our detailed experiments (cf. Machine Tree on the website `www.datagrapple.com`) and Python code available (`www.datagrapple.com/Tech`) for reproducible research. Finally, we conclude the paper with a discussion on the method and we propound future research directions.

Motivation and goal of study

Machine learning methodology usually consists in several pre-processing steps aiming at cleaning data and preparing them for being fed to a battery of algorithms. Data scientists have the daunting mission to choose the best possible combination of pre-processing, dissimilarity measure and algorithm to solve the task at hand among a profuse literature. In this article, we provide both a pre-processing and a distance for studying i.i.d. random processes which are compatible with basic machine learning algorithms.

Many statistical distances exist to measure the dissimilarity of two random variables, and therefore two i.i.d. random processes. Such distances can be roughly classified in two families:

1. distributional distances, for instance [184], [110] and [96], which focus on dissimilarity between probability distributions and quantify divergences in marginal behaviours,

2. dependence distances, such as the distance correlation or copula-based kernel dependency measures [172], which focus on the joint behaviours of random variables, generally ignoring their distribution properties.

However, we may want to be able to discriminate random variables both on distribution and dependence. This can be motivated, for instance, from the study of financial assets returns: are two perfectly correlated random variables (assets returns), but one being normally distributed and the other one following a heavy-tailed distribution, similar? From risk perspective, the answer is no [105], hence the propounded distance of this article. We illustrate its benefits through clustering, a machine learning task which primarily relies on the metric space considered (data representation and associated distance). Besides clustering has found application in finance, e.g. [207], which gives us a framework for benchmarking on real data.

Our objective is therefore to obtain a good clustering of random variables based on an appropriate and simple enough distance for being used with basic clustering algorithms, e.g. Ward hierarchical clustering [226], k-means++ [5], affinity propagation [84].

By clustering we mean the task of grouping sets of objects in such a way that objects in the same cluster are more similar to each other than those in different clusters. More specifically, a cluster of random variables should gather random variables with common dependence between them and with a common distribution. Two clusters should differ either in the dependency between their random variables or in their distributions.

A good clustering is a partition of the data that must be stable to small perturbations of the dataset. “Stability of some kind is clearly a desirable property of clustering methods" [49]. In the case of random variables, these small perturbations can be obtained from resampling [124], [151], [119] in the spirit of the bootstrap method since it preserves the statistical properties of the initial sample [75].

Yet, practitioners and researchers pinpoint that state-of-the-art results of clustering methodology applied to financial times series are very sensitive to perturbations [123]. The observed unstability may result from a poor representation of these time series, and thus clusters may not capture all the underlying information.

Shortcomings of a standard machine learning approach

A naive but often used distance between random variables to measure similarity and to perform clustering is the L_2 distance $\mathbb{E}[(X-Y)^2]$. Yet, this distance is not suited to our task.

Example 1 (Distance L_2 between two Gaussians)**.** *Let (X,Y) be a bivariate Gaussian vector, with $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$, $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ and whose correlation is $\rho(X,Y) \in [-1,1]$. We obtain $\mathbb{E}[(X-Y)^2] = (\mu_X - \mu_Y)^2 + (\sigma_X - \sigma_Y)^2 + 2\sigma_X\sigma_Y(1-\rho(X,Y))$. Now, consider the following values for correlation:*

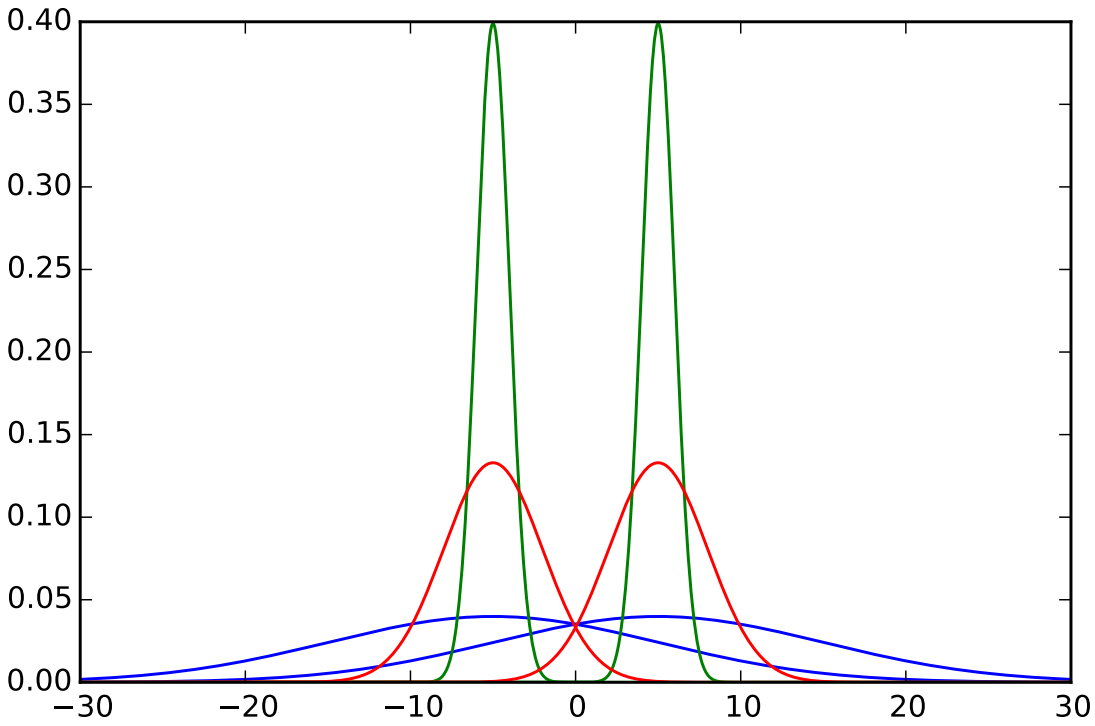


Figure 5.1: Probability density functions of Gaussians $\mathcal{N}(-5,1)$ and $\mathcal{N}(5,1)$ (in green), Gaussians $\mathcal{N}(-5,3)$ and $\mathcal{N}(5,3)$ (in red), and Gaussians $\mathcal{N}(-5,10)$ and $\mathcal{N}(5,10)$ (in blue). Green, red and blue Gaussians are equidistant using L_2 geometry on the parameter space (μ,σ).

- $\rho(X,Y)=0$, *so* $\mathbb{E}[(X-Y)^2]=(\mu_X-\mu_Y)^2+\sigma_X^2+\sigma_Y^2$. *The two variables are independent (since uncorrelated and jointly normally distributed), thus we must discriminate on distribution information. Assume* $\mu_X=\mu_Y$ *and* $\sigma_X=\sigma_Y$. *For* $\sigma_X=\sigma_Y\gg 1$, *we obtain* $\mathbb{E}[(X-Y)^2]\gg 1$ *instead of the distance* 0, *expected from comparing two equal Gaussians.*

- $\rho(X,Y)=1$, *so* $\mathbb{E}[(X-Y)^2]=(\mu_X-\mu_Y)^2+(\sigma_X-\sigma_Y)^2$. *Since the variables are perfectly correlated, we must discriminate on distributions. We actually compare them with a* L_2 *metric on the mean* $\times$ *standard deviation half-plane. However, this is not an appropriate geometry for comparing two Gaussians [55]. For instance, if* $\sigma_X=\sigma_Y=\sigma$, *we find* $\mathbb{E}[(X-Y)^2]=(\mu_X-\mu_Y)^2$ *for any values of* σ. *As* σ *grows, probability attached by the two Gaussians to a given interval grows similar (cf. Fig. 5.1), yet this increasing similarity is not taken into account by this* L_2 *distance.*

$\mathbb{E}[(X-Y)^2]$ considers both dependence and distribution information of the random variables, but not in a relevant way with respect to our task. Yet, we will benchmark against this distance since other more sophisticated distances on time series such as dynamic time warping [15] and representations such as wavelets [170] or SAX [129] were explicitly designed to handle temporal patterns which are inexistant in i.i.d. random processes.

A generic representation for random variables

Our purpose is to introduce a new data representation and a suitable distance which takes into account both distributional proximities and joint behaviours.

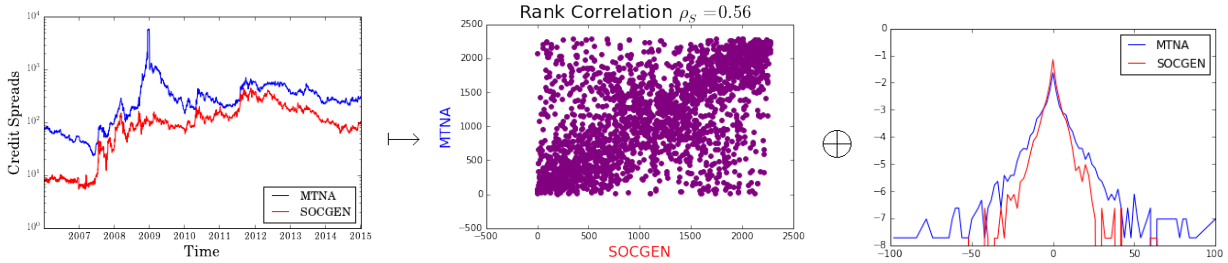


Figure 5.2: ArcelorMittal and Société générale prices (T observations $(X_1^t, X_2^t)_{t=1}^T$ from $(X_1, X_2) \in \mathcal{V}^2$) are projected on dependence $\oplus$ distribution space; $(G_{X_1}(X_1), G_{X_2}(X_2)) \in \mathcal{U}^2$ encode the dependence between X_1 and X_2 (a perfect correlation would be represented by a sharp diagonal on the scatterplot); (G_{X_1}, G_{X_2}) are the margins (their log-densities are displayed above), notice their heavy-tailed exponential distribution (especially for ArcelorMittal).

A representation preserving total information

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Ω is the sample space, $\mathcal{F}$ is the σ-algebra of events, and $\mathbb{P}$ is the probability measure. Let $\mathcal{V}$ be the space of all continuous real-valued random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$. Let $\mathcal{U}$ be the space of random variables following a uniform distribution on $[0, 1]$ and $\mathcal{G}$ be the space of absolutely continuous cumulative distribution functions (cdf).

Definition 4 (The copula transform)**.** *Let $X = (X_1, \dots, X_N) \in \mathcal{V}^N$ be a random vector with cdfs $G_X = (G_{X_1}, \dots, G_{X_N}) \in \mathcal{G}^N$. The random vector $G_X(X) = (G_{X_1}(X_1), \dots, G_{X_N}(X_N)) \in \mathcal{U}^N$ is known as the copula transform.*

Property 8 (Uniform margins of the copula transform)**.** *$G_{X_i}(X_i)$, $1 \leq i \leq N$, are uniformly distributed on $[0, 1]$.*

Proof. $x = G_{X_i}(G_{X_i}^{-1}(x)) = \mathbb{P}(X_i \leq G_{X_i}^{-1}(x)) = \mathbb{P}(G_{X_i}(X_i) \leq x)$. □

We define the following representation of random vectors that actually splits the joint behaviours of the marginal variables from their distributional information.

Definition 5 (dependence $\oplus$ distribution space projection)**.** *Let $\mathcal{T}$ be a mapping which transforms $X = (X_1, \dots, X_N)$ into its generic representation, an element of $\mathcal{U}^N \times \mathcal{G}^N$ representing X, defined as follow*

$$\begin{aligned} \mathcal{T} : \mathcal{V}^N &\to \mathcal{U}^N \times \mathcal{G}^N \\ X &\mapsto (G_X(X), G_X). \end{aligned} \tag{5.1}$$

Property 9. *$\mathcal{T}$ is a bijection.*

Proof. $\mathcal{T}$ is surjective as any element $(U, G) \in \mathcal{U}^N \times \mathcal{G}^N$ has the fiber $G^{-1}(U)$. $\mathcal{T}$ is injective as $(U_1, G_1) = (U_2, G_2)$ *a.s.* in $\mathcal{U}^N \times \mathcal{G}^N$ implies that they have the same cdf $G = G_1 = G_2$ and since $U_1 = U_2$ *a.s.*, it follows that $G^{-1}(U_1) = G^{-1}(U_2)$ *a.s.* □

This result replicates the seminal result of copula theory, namely Sklar's theorem [192], which asserts one can split the dependency and distribution apart without losing any information. Fig. 5.2 illustrates this projection for $N = 2$.

A distance between random variables

We leverage the propounded representation to build a suitable yet simple distance between random variables which is invariant under diffeomorphism.

Definition 6 (Distance d_θ between two random variables)**.** *Let $\theta \in [0,1]$. Let $(X,Y) \in \mathcal{V}^2$. Let $G = (G_X, G_Y)$, where G_X and G_Y are respectively X and Y marginal cdfs. We define the following distance*

$$d_\theta^2(X,Y) = \theta d_1^2(G_X(X), G_Y(Y)) + (1-\theta) d_0^2(G_X, G_Y), \tag{5.2}$$

where

$$d_1^2(G_X(X), G_Y(Y)) = 3\mathbb{E}[|G_X(X) - G_Y(Y)|^2], \tag{5.3}$$

and

$$d_0^2(G_X, G_Y) = \frac{1}{2}\int_{\mathbf{R}} \left(\sqrt{\frac{dG_X}{d\lambda}} - \sqrt{\frac{dG_Y}{d\lambda}} \right)^2 \mathrm{d}\lambda. \tag{5.4}$$

In particular, $d_0 = \sqrt{1-BC}$ is the Hellinger distance related to the Bhattacharyya (1/2-Chernoff) coefficient BC upper bounding the Bayes' classification error. To quantify distribution dissimilarity, d_0 is used rather than the more general α-Chernoff divergences since it satisfies the properties 10, 11, 12 (significant for practitioners). In addition, d_θ can thus be efficiently implemented as a scalar product. $d_1 = \sqrt{(1-\rho_S)/2}$ is a distance correlation measuring statistical dependence between two random variables, where ρ_S is the Spearman's correlation between X and Y. Notice that d_1 can be expressed by using the copula $C : [0,1]^2 \to [0,1]$ implicitly defined by the relation $G(X,Y) = C(G_X(X), G_Y(Y))$ since $\rho_S(X,Y) = 12\int_0^1\int_0^1 C(u,v)\,\mathrm{d}u\,\mathrm{d}v - 3$ [83].

Example 2 (Distance d_θ between two Gaussians)**.** *Let (X,Y) be a bivariate Gaussian vector, with $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$, $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ and $\rho(X,Y) = \rho$. We obtain,*

$$d_\theta^2(X,Y) = \theta\frac{1-\rho_S}{2} + (1-\theta)\left(1 - \sqrt{\frac{2\sigma_X\sigma_Y}{\sigma_X^2+\sigma_Y^2}} e^{-\frac{1}{4}\frac{(\mu_X-\mu_Y)^2}{\sigma_X^2+\sigma_Y^2}}\right).$$

Remember that for perfectly correlated Gaussians ($\rho = \rho_S = 1$), we want to discriminate on their distributions. We can observe that

- for $\sigma_X, \sigma_Y \to +\infty$, then $d_0(X,Y) \to 0$, it alleviates a main shortcoming of the basic L_2 distance which is diverging to $+\infty$ in this case;
- if $\mu_X \neq \mu_Y$, for $\sigma_X, \sigma_Y \to 0$, then $d_0(X,Y) \to 1$, its maximum value, i.e. it means that two Gaussians cannot be more remote from each other than two different Dirac delta functions.

We will refer to the use of this distance as the generic parametric representation (GPR) approach. GPR distance is a fast and good proxy for distance d_θ when the first two moments μ and σ predominate. Nonetheless, for datasets which contain heavy-tailed distributions, GPR fails to capture this information.

Property 10. *Let $\theta \in [0,1]$. The distance d_θ verifies $0 \leq d_\theta \leq 1$.*

Proof. Let $\theta \in [0,1]$. We have

1. $0 \leq d_0 \leq 1$, property of the Hellinger distance;
2. $0 \leq d_1 \leq 1$, since $-1 \leq \rho_S \leq 1$.

Finally, by convex combination, $0 \leq d_\theta \leq 1$. □

Property 11. *For $0 < \theta < 1$, d_θ is a metric.*

Proof. Let $(X, Y) \in \mathcal{V}^2$. For $0 < \theta < 1$, d_θ is a metric, and the only non-trivial property to verify is the separation axiom

1. $X = Y$ *a.s.* $\Rightarrow d_\theta(X, Y) = 0$
   $X = Y$ *a.s.* $\Rightarrow d_1(G_X(X), G_Y(Y)) = d_0(G_X, G_Y) = 0$, and thus $d_\theta(X, Y) = 0$,

2. $d_\theta(X, Y) = 0 \Rightarrow X = Y$ *a.s.*
   $d_\theta(X, Y) = 0 \Rightarrow d_1(G_X(X), G_Y(Y)) = 0$ and $d_0(G_X, G_Y) = 0 \Rightarrow G_X(X) = G_Y(Y)$ *a.s.* and $G_X = G_Y$. Since G is absolutely continuous, it follows $X = Y$ *a.s.*

Notice that for $\theta \in \{0, 1\}$, this property does not hold. Let $U \in \mathcal{V}$, $U \sim \mathcal{U}[0, 1]$. $U \neq 1 - U$ but $d_0(U, 1 - U) = 0$. Let $V \in \mathcal{V}$. $V \neq 2V$ but $d_1(V, 2V) = 0$. □

Property 12. *Diffeomorphism invariance. Let $h : \mathcal{V} \to \mathcal{V}$ be a diffeomorphism. Let $(X, Y) \in \mathcal{V}^2$. Distance d_θ is invariant under diffeomorphism, i.e.*

$$d_\theta(h(X), h(Y)) = d_\theta(X, Y). \tag{5.5}$$

Proof. From definition, we have

$$d_0^2(h(X), h(Y)) = 1 - \int_{\mathbf{R}} \sqrt{\frac{dG_{h(X)}}{d\lambda} \frac{dG_{h(Y)}}{d\lambda}} d\lambda \tag{5.6}$$

and since

$$\frac{dG_{h(X)}}{d\lambda}(\lambda) = \frac{1}{h'\left(h^{-1}(\lambda)\right)} \frac{dG_X}{d\lambda}\left(h^{-1}(\lambda)\right), \tag{5.7}$$

we obtain

$$\begin{aligned} d_0^2(h(X), h(Y)) &= 1 - \int_{\mathbf{R}} \frac{1}{h'\left(h^{-1}(\lambda)\right)} \sqrt{\frac{dG_X}{d\lambda} \frac{dG_Y}{d\lambda}}\left(h^{-1}(\lambda)\right) d\lambda \\ &= d_0^2(X, Y). \end{aligned} \tag{5.8}$$

In addition, $\forall x \in \mathbf{R}$, we have

$$\begin{aligned} G_{h(X)}\left(h(x)\right) &= \mathbb{P}\left[h(X) \leq h(x)\right] \\ &= \begin{cases} \mathbb{P}\left[X \leq x\right] = G_X(x), \text{ if } h \text{ increasing} \\ 1 - \mathbb{P}\left[X \leq x\right] = 1 - G_X(x), \text{ otherwise} \end{cases} \end{aligned} \tag{5.9}$$

which implies that

$$\begin{aligned} d_1^2\left(h(X), h(Y)\right) &= 3\mathbb{E}\left[\left|G_{h(X)}(h(X)) - G_{h(Y)}(h(Y))\right|^2\right] \\ &= 3\mathbb{E}\left[\left|G_X(X) - G_Y(Y)\right|^2\right] \\ &= d_1^2(X, Y). \end{aligned} \tag{5.10}$$

Finally, we obtain Property 12 by definition of d_θ. □

Thus, d_θ is invariant under monotonic transformations, a desirable property as it ensures to be insensitive to scaling (e.g. choice of units) or measurement scheme (e.g. device, mathematical modelling) of the underlying phenomenon.

A non-parametric statistical estimation of d_θ

To apply the propounded distance d_θ on sampled data without parametric assumptions, we have to define its statistical estimate $\tilde{d}_\theta$ working on realizations of the i.i.d. random variables. Distance d_1 working with continuous uniform distributions can be approximated by normalized rank statistics yielding to discrete uniform distributions, in fact coordinates of the multivariate empirical copula [62] which is a non-parametric estimate converging uniformly toward the underlying copula [61]. Distance d_0 working with densities can be approximated by using its discrete form working on histogram density estimates.

Definition 7 (The empirical copula transform)**.** *Let $X^T = (X_1^t, \ldots, X_N^t)$, $t = 1, \ldots, T$, be T observations from a random vector $X = (X_1, \ldots, X_N)$ with continuous margins $G_X = (G_{X_1}(X_1), \ldots, G_{X_N}(X_N))$. Since one cannot directly obtain the corresponding copula observations $(G_{X_1}(X_1^t), \ldots, G_{X_N}(X_N^t))$ without knowing a priori G_X, one can instead estimate the N empirical margins $G_{X_i}^T(x) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}(X_i^t \leq x)$ to obtain T empirical observations $(G_{X_1}^T(X_1^t), \ldots, G_{X_N}^T(X_N^t))$ which are thus related to normalized rank statistics as $G_{X_i}^T(X_i^t) = X_i^{(t)}/T$, where $X_i^{(t)}$ denotes the rank of observation X_i^t.*

Definition 8 (Empirical distance)**.** *Let $(X^t)_{t=1}^T$ and $(Y^t)_{t=1}^T$ be T realizations of real-valued random variables $X, Y \in \mathcal{V}$ respectively. An empirical distance between realizations of random variables can be defined by*

$$\tilde{d}_\theta^2\left((X^t)_{t=1}^T, (Y^t)_{t=1}^T\right) \overset{a.s.}{=} \theta\tilde{d}_1^2 + (1-\theta)\tilde{d}_0^2, \tag{5.11}$$

where

$$\tilde{d}_1^2 = \frac{3}{T(T^2-1)}\sum_{t=1}^{T}\left(X^{(t)} - Y^{(t)}\right)^2 \tag{5.12}$$

and

$$\tilde{d}_0^2 = \frac{1}{2}\sum_{k=-\infty}^{+\infty}\left(\sqrt{g_X^h(hk)} - \sqrt{g_Y^h(hk)}\right)^2, \tag{5.13}$$

h being here a suitable bandwidth, and $g_X^h(x) = \frac{1}{T}\sum_{t=1}^{T}\mathbf{1}(\lfloor\frac{x}{h}\rfloor h \leq X^t < (\lfloor\frac{x}{h}\rfloor + 1)h)$ being a density histogram estimating pdf g_X from $(X^t)_{t=1}^T$, T realizations of random variable $X \in \mathcal{V}$.

We will refer henceforth to this distance and its use as the generic non-parametric representation (GNPR) approach. To use effectively d_θ and its statistical estimate, it boils down to select a particular value for θ. We suggest here an exploratory approach where one can test (i) distribution information ($\theta = 0$), (ii) dependence information ($\theta = 1$), and (iii) a mix of both information ($\theta = 0.5$). Ideally, θ should reflect the balance of dependence and distribution information in the data. In a supervised setting, one could select an estimate $\hat{\theta}$ of the right balance $\theta^\star$ optimizing some loss function by techniques such as cross-validation. Yet, the lack of a clear loss function makes the estimation of $\theta^\star$ difficult in an unsupervised setting. For clustering, many authors [119], [189], [190], [147] suggest stability as a tool for parameter selection. But, [14] warn against its irrelevant use for this purpose. Besides, we already use stability for clustering validation and we want to avoid overfitting. Finally, we think that finding an optimal trade-off $\theta^\star$ is important for accelerating the rate of convergence toward the underlying ground truth when working with finite and possibly small samples, but ultimately lose its importance asymptotically as soon as $0 < \theta < 1$.

Experiments and applications

Synthetic datasets

We propose the following model for testing the efficiency of the GNPR approach: N time series of length T which are subdivided into K correlation clusters themselves subdivided into D distribution clusters.

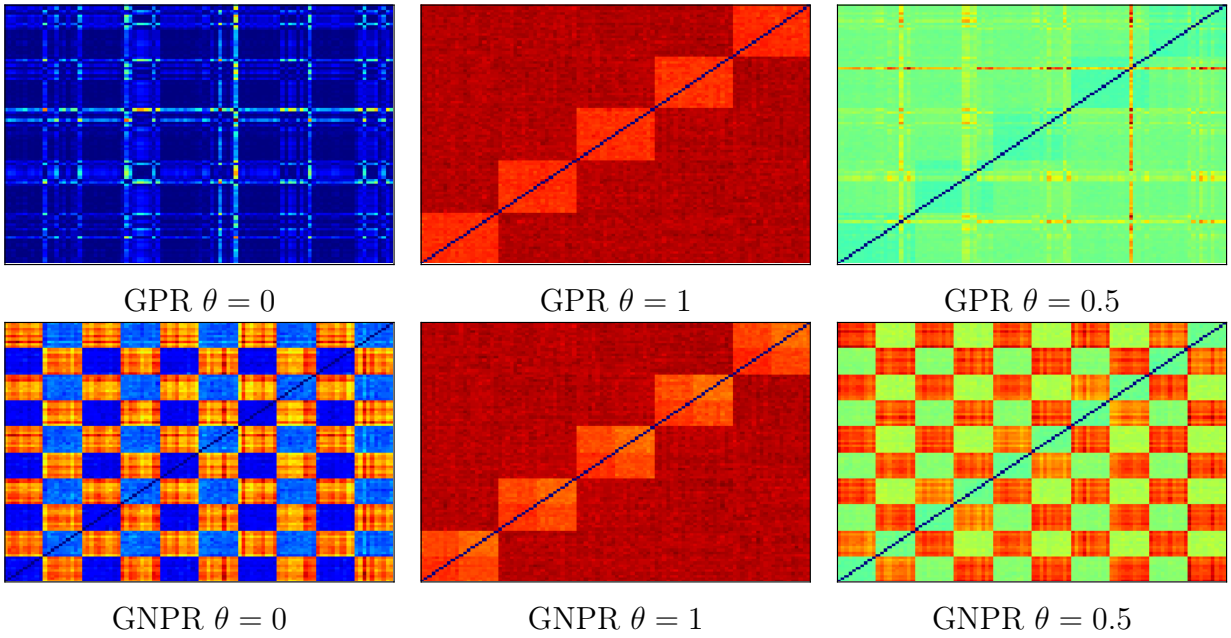


Figure 5.3: GPR and GNPR distance matrices. Both GPR and GNPR highlight the 5 correlation clusters ($\theta = 1$), but only GNPR finds the 2 distributions ($\mathcal{S}$ and $\mathcal{L}$) subdividing them ($\theta = 0$). Finally, by combining both information GNPR ($\theta = 0.5$) can highlight the 10 original clusters, while GPR ($\theta = 0.5$) simply adds noise on the correlation distance matrix it recovers.

Table 5.1: Model parameters for some interesting test case datasets

Clustering	Dataset	N	T	Q	K	β	Y_k	Z_1^i	Z_2^i	Z_3^i	Z_4^i
Distribution	A	200	5000	4	1	0	$\mathcal{N}(0,1)$	$\mathcal{N}(0,1)$	$\mathcal{L}$	$\mathcal{S}$	$\mathcal{N}(0,2)$
Dependence	B	200	5000	10	10	0.1	$\mathcal{S}$	$\mathcal{S}$	$\mathcal{S}$	$\mathcal{S}$	$\mathcal{S}$
Mix	C	200	5000	10	5	0.1	$\mathcal{N}(0,1)$	$\mathcal{N}(0,1)$	$\mathcal{S}$	$\mathcal{N}(0,1)$	$\mathcal{S}$
	G	$32, \ldots, 640$	$10, \ldots, 2000$	32	8	0.1	$\mathcal{N}(0,1)$	$\mathcal{N}(0,1)$	$\mathcal{N}(0,2)$	$\mathcal{L}$	$\mathcal{S}$

Let $(Y_k)_{k=1}^K$, be K i.i.d. random variables. Let $p, D \in \mathbf{N}$. Let $N = pKD$. Let $(Z_d^i)_{d=1}^D$, $1 \leq i \leq N$, be independent random variables. For $1 \leq i \leq N$, we define

$$X_i = \sum_{k=1}^{K} \beta_{k,i} Y_k + \sum_{d=1}^{D} \alpha_{d,i} Z_d^i, \tag{5.14}$$

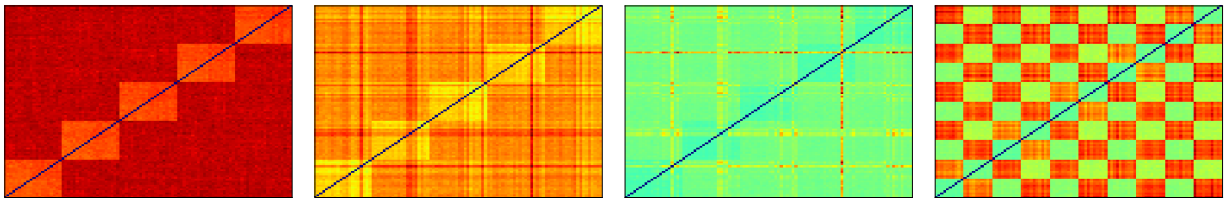
where

1. $\alpha_{d,i} = 1$, if $i \equiv d - 1 \pmod{D}$, 0 otherwise;
2. $\beta \in [0, 1]$,
3. $\beta_{k,i} = \beta$, if $\lceil iK/N \rceil = k$, 0 otherwise.

$(X_i)_{i=1}^N$ are partitioned into $Q = KD$ clusters of p random variables each. Playing with the model parameters, we define in Table 5.1 some interesting test case datasets to study distribution clustering, dependence clustering or a mix of both. We use the following notations as a shorthand: $\mathcal{L} := \text{Laplace}(0, 1/\sqrt{2})$ and $\mathcal{S} := \text{t-distribution}(3)/\sqrt{3}$. Since $\mathcal{L}$ and $\mathcal{S}$ have both a mean of 0 and a variance of 1, GPR cannot find any difference between them, but GNPR can discriminate on higher moments as it can be seen in Fig. 5.3.

Performance of clustering using GNPR

We empirically show that the GNPR approach achieves better results than others using common distances regardless of the algorithm used on the defined test cases A, B and C described in Table 5.1. Test case A illustrates datasets containing only distribution information: there are 4 clusters of distributions. Test case B illustrates datasets containing



$(1-\rho)/2$ L_2 GPR $\theta = 0.5$ GNPR $\theta = 0.5$

Figure 5.4: Distance matrices obtained on dataset C using distance correlation, L_2 distance, GPR and GNPR. None but GNPR highlights the 10 original clusters which appear on its diagonal.

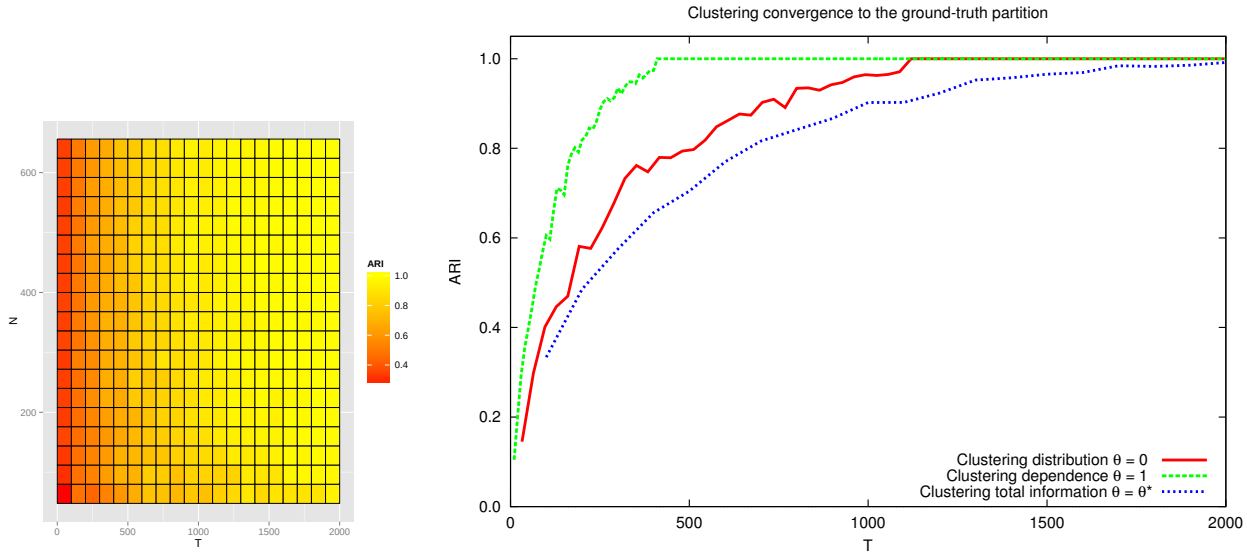


Figure 5.5: Empirical consistency of clustering using GNPR as $T \to \infty$

only dependence information: there are 10 clusters of correlation between random variables which are heavy-tailed. Test case C illustrates datasets containing both information: it consists in 10 clusters composed of 5 correlation clusters and each of them is divided into 2 distribution clusters. Using scikit-learn implementation [167], we apply 3 clustering algorithms with different paradigms: a hierarchical clustering using average linkage (HC-AL), k-means++ (KM++), and affinity propagation (AP). Experiment results are reported in Table 5.2. GNPR performance is due to its proper representation (cf. Fig. 5.4). Finally, we have noticed increasing precision of clustering using GNPR as time T grows to infinity, all other parameters being fixed. The number of time series N seems rather uninformative as illustrated in Fig. 5.5 (left) which plots ARI [100] between computed clustering and ground-truth of dataset G as an heatmap for varying N and T. Fig. 5.5 (right) shows the convergence to the true clustering as a function of T.

Application to financial time series clustering

Clustering assets: a (too) strong focus on correlation

It has been noticed that straightfoward approaches automatically discover sector and industries [133]. Since detected patterns are blatantly correlation-flavoured, it prompted econophysicists to focus on correlations, hierarchies and networks [212] from the Minimum Spanning Tree and its associated clustering algorithm the Single Linkage to the state of the art [155] exploiting the topological properties of the Planar Maximally Filtered Graph [214] and its associated algorithm the Directed Bubble Hierarchical Tree (DBHT) technique [197]. In practice, econophysicists consider the assets log returns and compute their correlation matrix. The correlation matrix is then filtered thanks to a clustering of the correlation-network

Table 5.2: Comparison of distance correlation, L_2 distance, GPR and GNPR: GNPR approach improves clustering performance

		Adjusted Rand Index		
Algo.	Distance	A	B	C
HC-AL	$(1-\rho)/2$	0.00 $\pm$0.01	0.99 $\pm$0.01	0.56 $\pm$0.01
	$\mathbb{E}[(X-Y)^2]$	0.00 $\pm$0.00	0.09 $\pm$0.12	0.55 $\pm$0.05
	GPR $\theta=0$	0.34 $\pm$0.01	0.01 $\pm$0.01	0.06 $\pm$0.02
	GPR $\theta=1$	0.00 $\pm$0.01	0.99 $\pm$0.01	0.56 $\pm$0.01
	GPR $\theta=.5$	0.34 $\pm$0.01	0.59 $\pm$0.12	0.57 $\pm$0.01
	GNPR $\theta=0$	**1**	0.00 $\pm$0.00	0.17 $\pm$0.00
	GNPR $\theta=1$	0.00 $\pm$0.00	**1**	0.57 $\pm$0.00
	GNPR $\theta=.5$	0.99 $\pm$0.01	0.25 $\pm$0.20	**0.95** $\pm$0.08
KM++	$(1-\rho)/2$	0.00 $\pm$0.01	0.60 $\pm$0.20	0.46 $\pm$0.05
	$\mathbb{E}[(X-Y)^2]$	0.00 $\pm$0.00	0.34 $\pm$0.11	0.48 $\pm$0.09
	GPR $\theta=0$	0.41 $\pm$0.03	0.01 $\pm$0.01	0.06 $\pm$0.02
	GPR $\theta=1$	0.00 $\pm$0.00	0.45 $\pm$0.11	0.43 $\pm$0.09
	GPR $\theta=.5$	0.27 $\pm$0.05	0.51 $\pm$0.14	0.48 $\pm$0.06
	GNPR $\theta=0$	**0.96** $\pm$0.11	0.00 $\pm$0.01	0.14 $\pm$0.02
	GNPR $\theta=1$	0.00 $\pm$0.01	**0.65** $\pm$0.13	0.53 $\pm$0.02
	GNPR $\theta=.5$	0.72 $\pm$0.13	0.21 $\pm$0.07	**0.64** $\pm$0.10
AP	$(1-\rho)/2$	0.00 $\pm$0.00	0.99 $\pm$0.07	0.48 $\pm$0.02
	$\mathbb{E}[(X-Y)^2]$	0.14 $\pm$0.03	0.94 $\pm$0.02	0.59 $\pm$0.00
	GPR $\theta=0$	0.25 $\pm$0.08	0.01 $\pm$0.01	0.05 $\pm$0.02
	GPR $\theta=1$	0.00 $\pm$0.01	0.99 $\pm$0.01	0.48 $\pm$0.02
	GPR $\theta=.5$	0.06 $\pm$0.00	0.80 $\pm$0.10	0.52 $\pm$0.02
	GNPR $\theta=0$	**1**	0.00 $\pm$0.00	0.18 $\pm$0.01
	GNPR $\theta=1$	0.00 $\pm$0.01	**1**	0.59 $\pm$0.00
	GNPR $\theta=.5$	0.39 $\pm$0.02	0.39 $\pm$0.11	**1**

[63] built using similarity and dissimilarity matrices which are derived from the correlation one by convenient *ad hoc* transformations. Clustering these correlation-based networks [160] aims at filtering the correlation matrix for standard portfolio optimization [207]. Yet, adopting similar approaches only allow to retrieve information given by assets co-movements and nothing about the specificities of their returns behaviour, whereas we may also want to distinguish assets by their returns distribution. For example, we are interested to know whether they undergo fat tails, and to which extent.

Clustering credit default swaps

We apply the GNPR approach on financial time series, namely daily credit default swap [101] (CDS) prices. We consider the $N = 500$ most actively traded CDS according to DTCC (`http://www.dtcc.com/`). For each CDS, we have $T = 2300$ observations corresponding to historical daily prices over the last 9 years, amounting for more than one million data points. Since credit default swaps are traded over-the-counter, closing time for fixing prices can be arbitrarily chosen, here 5pm GMT, i.e. after the London Stock Exchange trading session. This synchronous fixing of CDS prices avoids spurious correlations arising from different closing times. For example, the use of close-to-close stock prices artificially overestimates intra-market correlation and underestimates inter-market dependence since they have different trading hours [135]. These CDS time series can be consulted on the web portal `http://www.datagrapple.com/`.

Assuming that CDS prices $(P^t)_{t\geq 1}$ follow random walks, their increments $\Delta P^t = P^t - P^{t-1}$

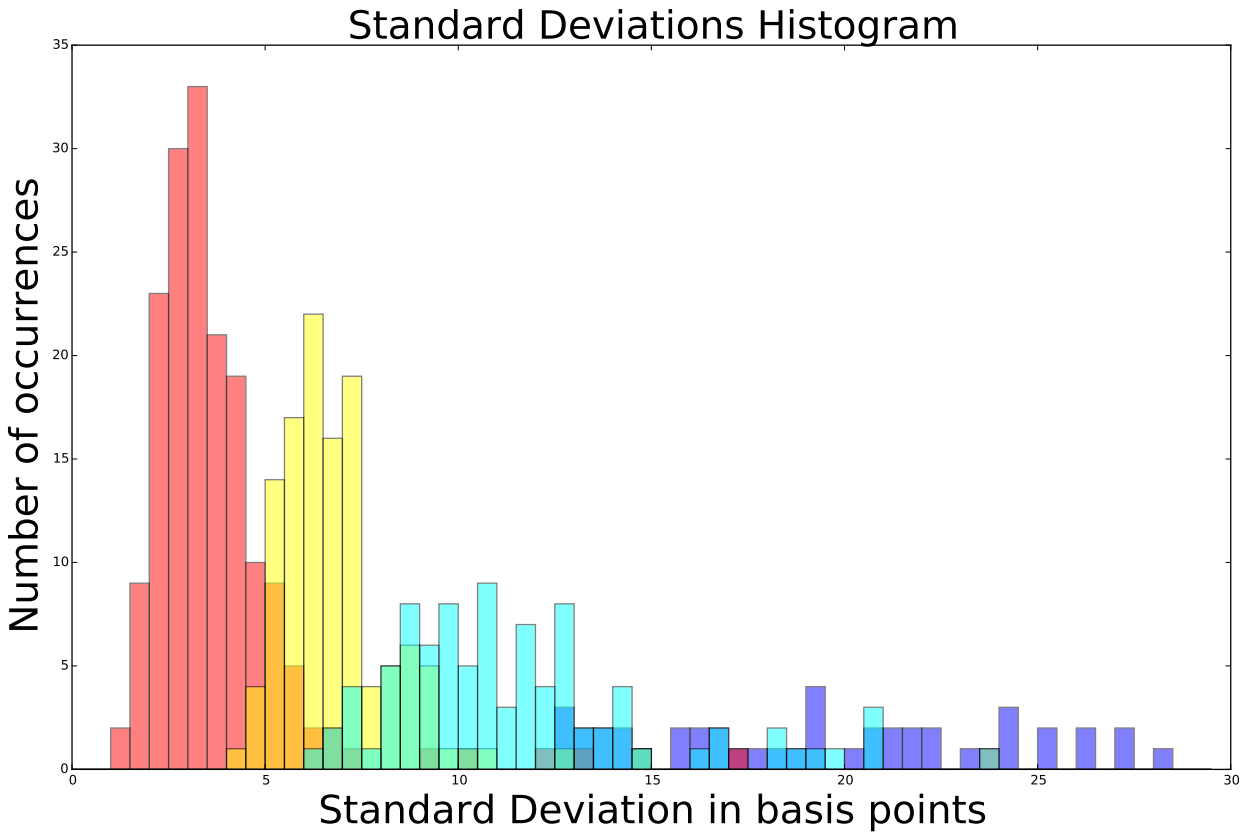


Figure 5.6: Standard Deviation Histogram. The 4 clusters found using GNPR $\theta = 0$ represented by the 4 colors fit precisely the multi-modal distribution of standard deviations.

are i.i.d. random variables, and therefore the GNPR approach can be applied to the time series of prices variations, i.e. on data $(\Delta P_1^t, \ldots, \Delta P_N^t)$, $t = 1, \ldots, T$. Thus, for aggregating CDS prices time series, we use a clustering algorithm (for instance, Ward's method [226]) based on the GNPR distance matrices between their variations.

Using GNPR $\theta = 0$, we look for distribution information in our CDS dataset. We observe that clustering based on the GNPR $\theta = 0$ distance matrix yields 4 clusters which fit precisely the multi-modal empirical distribution of standard deviations as can be seen in Fig. 5.6. For GNPR $\theta = 1$, we display in Fig. 5.7 the rank correlation distance matrix obtained. We can notice its hierarchical structure already described in many papers, e.g. [133], [28], focusing on stock markets. There is information in distribution and in correlation, thus taking into account both information, i.e. using GNPR $\theta = 0.5$, should lead to a meaningful clustering. We verify this claim by using stability as a criterion for validation. Practically, we consider even and odd trading days and perform two independent clusterings, one on even days and the other one on odd days. We should obtain the same partitions. In Fig. 5.8, we display the partitions obtained using the GNPR $\theta = 0.5$ approach next to the ones obtained by applying a L_2 distance on prices returns. We find that GNPR clustering is more stable than L_2 on returns clustering. Moreover, clusters obtained from GNPR are more homogeneous in size.

To conclude on the experiments, we have highlighted through clustering that the presented approach leveraging dependence and distribution information leads to better results: finer partitions on synthetic test cases and more stable partitions on financial time series.

Discussion

In this paper, we have exposed a novel representation of random variables which could lead to improvements in applying machine learning techniques on time series describing underlying i.i.d. stochastic processes. We have empirically shown its relevance to deal with random walks and financial time series. We have led a large scale experiment on the credit derivatives

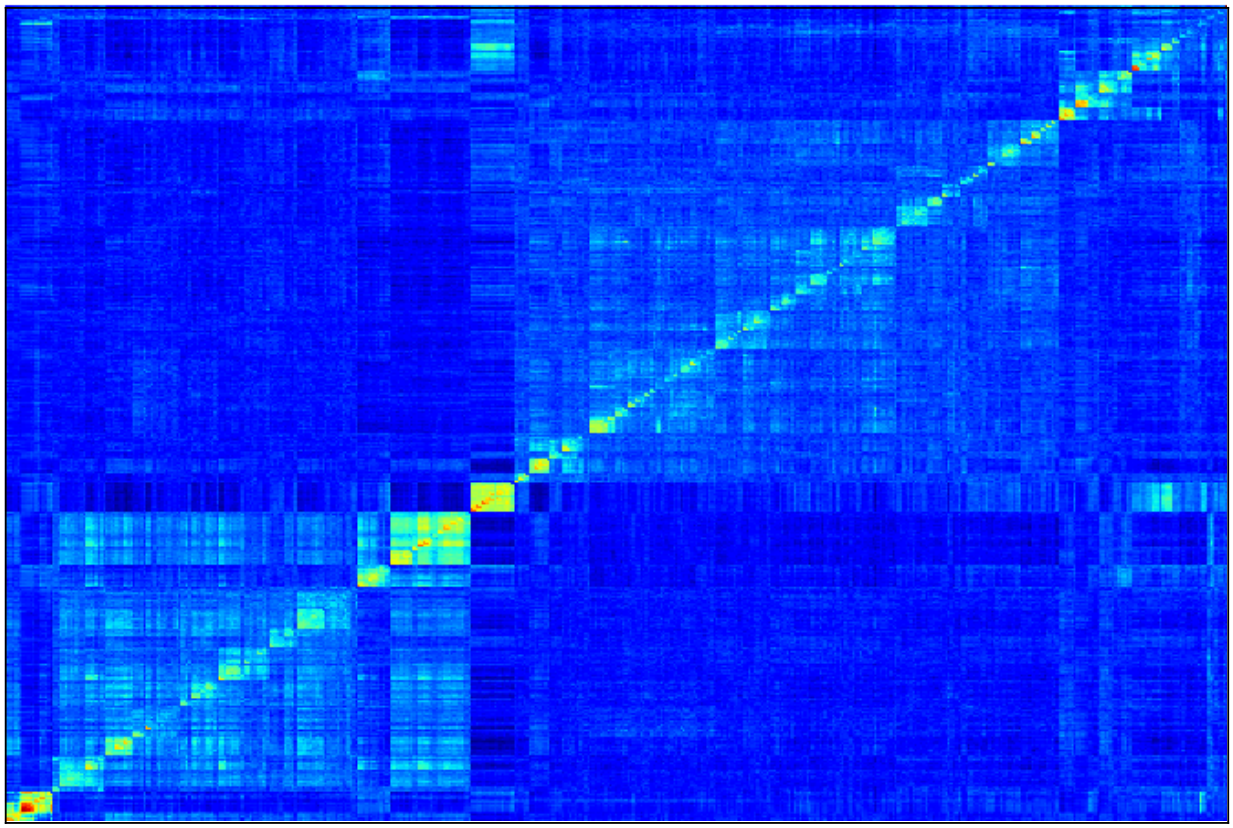


Figure 5.7: Centered Rank Correlation Distance Matrix. GNPR $\theta = 1$ exhibits a hierarchical structure of correlations: first level consists in Europe, Japan and US; second level corresponds to credit quality (investment grade or high yield); third level to industrial sectors.

market notorious for not having Gaussian but heavy-tailed returns, first results are available on website `www.datagrapple.com`. We also intend to lead such clustering experiments for testing applicability of the method to areas outside finance. On the theoretical side, we plan to improve the aggregation of the correlation and distribution part by using elements of information geometry theory and to study the consistency property of our method.

5.2 Alternatives to standard correlations

We propose a methodology to explore and measure the pairwise correlations that exist between variables in a dataset. The methodology leverages copulas for encoding dependence between two variables, state-of-the-art optimal transport for providing a relevant geometry to the copulas, and clustering for summarizing the main dependence patterns found between the variables. Some of the clusters centers can be used to parameterize a novel dependence coefficient which can target or forget specific dependence patterns. Finally, we illustrate and benchmark the methodology on several datasets. Code and numerical experiments are available online for reproducible research.

Introduction

Pearson's correlation coefficient which estimates linear dependence between two variables is still the mainstream tool for measuring variable correlations in science and engineering. However, its shortcomings are well-documented in the statistics literature: not robust to outliers; not invariant to monotone transformations of the variables; can take value 0 whereas variables are strongly dependent; only relevant when variables are jointly normally distributed. A large but under-exploited literature in statistics and machine learning has expanded recently to alleviate these issues [179, 201, 187]. An underlying idea to many of the dependence

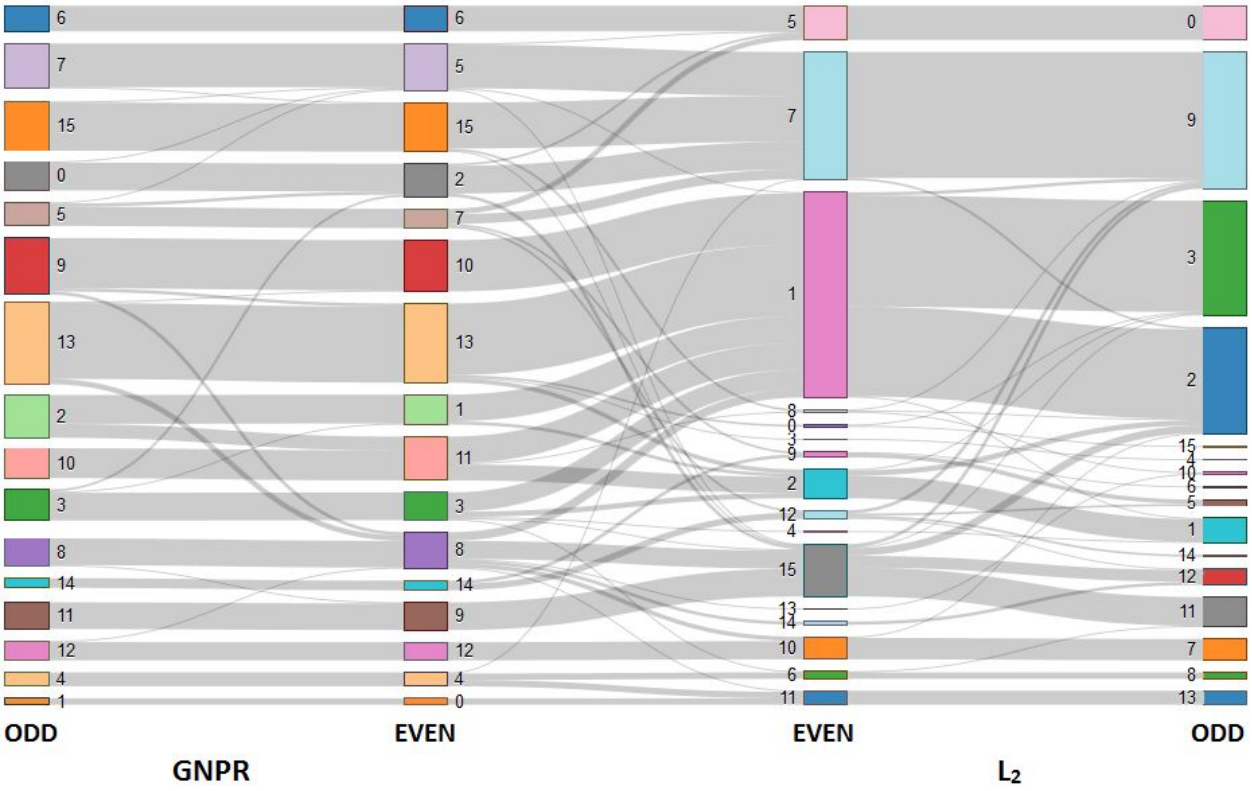


Figure 5.8: Better clustering stability using the GNPR approach: GNPR $\theta = 0.5$ achieves ARI $= 0.85$; L_2 on returns achieves ARI 0.64; The two leftmost partitions built from GNPR on the odd/even trading days sampling look similar: only a few CDS are switching from clusters; The two rightmost partitions built using a L_2 on returns display very inhomogeneous (odd-2,3,9 vs. odd-4,14,15) and unstable (even-1 splitting into odd-3 and odd-2) clusters.

coefficients is to compute a distance $D(P(X,Y), P(X)P(Y))$ between the joint distribution $P(X,Y)$ of variables X, Y and $P(X)P(Y)$ the product of marginal distributions encoding the independence. For example, choosing $D = \mathrm{KL}$ (Kullback-Leibler divergence), we end up with the Mutual Information (MI) measure, well-known in information theory. Thus, one can detect all the dependences between X and Y since the distance will be greater than 0 as soon as $P(X,Y)$ is different from $P(X)P(Y)$. Then, the dependence literature focus has shifted toward the new concept of "equitability" [113]: How can one quantify the strength of a statistical association between two variables without bias for relationships of a specific form? Many researchers now aim at designing and proving that their proposed measures are indeed equitable [178, 65, 50]. This is *not* what we look for in this article. But, on the contrary, we want to target specific dependence patterns and ignore others. We want to target dependence which are relevant to such or such problem, and forget about the dependence which are not in the scope of the problems at hand, or even worse which may be spurious associations (pure chance or artifacts in the data). The latter will be detected with an equitable dependence measure since they are deviation from independence, and will be given as much weight as the interesting ones. Rather than using the biases for specific dependence of several coefficients, we propose a dependence coefficient that can be parameterized by a set of *target-dependences*, and a set of *forget-dependences*. Sets of target and forget dependences can be built using expert hypotheses, or by leveraging the centers of clusters resulting from an exploratory clustering of the pairwise dependences. To achieve this goal, we will leverage three tools: copulas, optimal transportation, and clustering. Whereas clustering, the task of grouping a set of objects in such a way that objects in the same group (also called cluster) are more similar to each other than those in different groups, is common

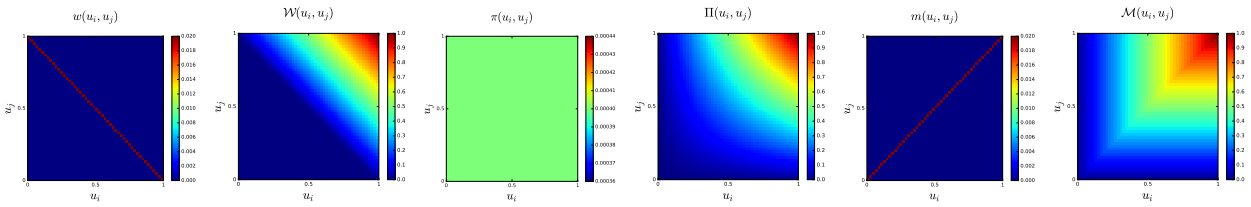


Figure 5.9: Copulas measure (left column) and cumulative distribution function (right column) heatmaps for negative dependence (first row), independence (second row), i.e. the uniform distribution over $[0,1]^2$, and positive dependence (third row)

knowledge in the machine learning community, copulas and optimal transportation are not yet mainstream tools. Copulas have recently gained attention in machine learning [77], and several copula-based dependence measures have been proposed for improving feature selection methods [88, 132, 50]. Optimal transport may be more familiar to computer scientists working in computer vision since it is the underlying theory of the Earth Mover's Distance [181]. Until very recently, optimal transportation distances between distributions were not deemed relevant for machine learning applications since the best computational cost known was super-cubic to the number of bins used for discretizing the distribution supports which grows itself exponentially with the dimension. A mere distance evaluation could take several seconds! In this article, we leverage recent computational breakthroughs detailed in [59] which make their use practical in machine learning.

Background on Copulas and Optimal Transport

Copulas

Copulas are functions that couple multivariate distribution functions to their one-dimensional marginal distribution functions [156]. In this article, we will only consider bivariate copulas, but most of the results and the methodology presented hold in the multivariate setting, at the cost of a much higher computational burden which is for now a bit unrealistic.

Definition 9 (Sklar's Theorem [193])**.** *For any random vector $X = (X_i, X_j)$ having continuous marginal cumulative distribution functions F_i, F_j respectively, its joint cumulative distribution F is uniquely expressed as $F(X_i, X_j) = C(F_i(X_i), F_j(X_j))$, where C, the bivariate distribution of uniform marginals $U_i, U_j := F_i(X_i), F_j(X_j)$, is known as the copula of X.*

Copulas are central for studying the dependence between random variables: their uniform marginals jointly encode all the dependence. They allow to study scale-free measures of dependence and are *invariant to monotonous transformations of the variables*. Some copulas play a major role in the measure of dependence, namely $\mathcal{W}$ and $\mathcal{M}$ the Fréchet-Hoeffding copula bounds, and the independence copula $\Pi(u_i, u_j) = u_i u_j$ (depicted in Figure 5.9).

Definition 10 (Fréchet-Hoeffding copula bounds)**.** *For any copula $C : [0,1]^2 \to [0,1]$ and any $(u_i, u_j) \in [0,1]^2$ the following bounds hold:*

$$\mathcal{W}(u_i, u_j) \leq C(u_i, u_j) \leq \mathcal{M}(u_i, u_j), \tag{5.15}$$

where $\mathcal{W}(u_i, u_j) = \max\{u_i + u_j - 1, 0\}$ is the copula for countermonotonic random variables and $\mathcal{M}(u_i, u_j) = \min\{u_i, u_j\}$ is the copula for comonotonic random variables.

Many correlation coefficients can actually be expressed as a distance between the data copula and one of these reference copulas. For example, the Spearman (rank) correlation ρ_S which is usually understood as $\rho_S(X_i, X_j) = \rho(F_i(X_i), F_j(X_j))$, i.e. the linear dependence of the probability integral transformed variables (rank-transformed data), can

also be viewed as an average distance between the copula C of (X_i, X_j) and the independence copula Π: $\rho_S(X_i, X_j) = 12 \int \int_{[0,1]^2} (C(u_i, u_j) - u_i u_j)\, du_i du_j$ [156]. Moreover, since $|u_i - u_j|/\sqrt{2}$ is the distance between point (u_i, u_j) to the diagonal (the measure of the positive dependence copula), one can rewrite $\rho_S(X_i, X_j) = 12 \int \int_{[0,1]^2} (C(u_i, u_j) - u_i u_j)\, du_i du_j = 12 \int \int_{[0,1]^2} u_i u_j dC(u_i, u_j) - 3 = 1 - 6 \int \int_{[0,1]^2} (u_i - u_j)^2 dC(u_i, u_j)$ [127]. Thus, Spearman correlation can also be viewed as measuring a deviation from the monotonically increasing dependence to the data copula using a quadratic distance. *We will leverage this idea to propose our dependence-parameterized dependence coefficient.*

Notice that when working with empirical data, we do not know a priori the margins F_i for applying the probability integral transform $U_i := F_i(X_i)$. Deheuvels in [62] has introduced a practical estimator for the uniform margins and the underlying copula, the empirical copula transform.

Definition 11 (Empirical Copula Transform)**.** *Let (X_i^t, X_j^t), $t = 1, \ldots, T$, be T observations from a random vector (X_i, X_j) with continuous margins. Since one cannot directly obtain the corresponding copula observations $(U_i^t, U_j^t) := (F_i(X_i^t), F_j(X_j^t))$, where $t = 1, \ldots, T$, without knowing a priori F_i, one can instead estimate the empirical margins $F_i^T(x) = \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}(X_i^t \leq x)$, to obtain the T empirical observations $(\tilde{U_i^t}, \tilde{U_j^t}) := (F_i^T(X_i^t), F_j^T(X_j^t))$. Equivalently, since $\tilde{U_i^t} = R_i^t/T$, R_i^t being the rank of observation X_i^t, the empirical copula transform can be considered as the normalized rank transform.*

Notice that the empirical copula transform is fast to compute, sorting arrays of length T can be done in $O(T \log T)$, consistent and converges fast to the underlying copula [61], [88].

As motivated in the introduction, we want to compare and summarize the pairwise empirical dependence structure (empirical bivariate copulas) of many variables. This brings the following questions: How can we compare two such copulas? What is a relevant representative of a set of empirical copulas? Which geometries are relevant for clustering these empirical distributions, and which are not?

Optimal Transport vs. Fisher-Rao Geometry of Gaussian copulas

Our approach (depicted in Fig. 5.2) is to leverage distances from Information Geometry to compare distributions - the copulas encoding dependence between the variates - in order to discriminate on the dependence between the random variables (but not on their distributions). What kind of distances is relevant for comparing copulas? Far from being comprehensive, we illustrate our point with Wasserstein distances, Fisher-Rao geodesic distance and related divergences.

Sensitivity of distances with respect to dependence

A reminder on statistical distances

Statistical distances are distances between probability distributions. Many such distances have been designed to deal with practical problems in signal processing [157].

One of the leading approaches is to consider the parameter space $\Theta = \{\theta \in \mathbf{R}^D \mid \int p_\theta(x)dx = 1\}$ of a family of parametric probability distributions $\{p_\theta(x)\}_\theta$ with $x \in \mathbf{R}^d$ and $\theta \in \mathbf{R}^D$ as a Riemannian manifold endowed with the Fisher-Rao metric $ds^2(\theta) = \sum_{i=1}^{D} \sum_{j=1}^{D} g_{ij}(\theta) d\theta_i d\theta_j$ [176]. The coefficients $g_{ij}(\theta) = \mathbf{E}_\theta \left[\frac{1}{p(\theta)} \frac{\partial p}{\partial \theta_i} \frac{1}{p(\theta)} \frac{\partial p}{\partial \theta_j} \right] = g_{ji}(\theta)$ are known as the Fisher Information Matrix coefficients. Two probability distributions represented by their respective density p_{θ_1} and p_{θ_2} are considered as two points θ_1 and θ_2 on the manifold (Θ, ds^2). The Fisher-Rao geodesic distance between these two probability distributions can be computed by integrating the Fisher-Rao metric along the geodesics (locally shortest paths) between the corresponding points θ_1 and θ_2: $D(\theta_1, \theta_2) = \int_{\theta_1}^{\theta_2} ds$.

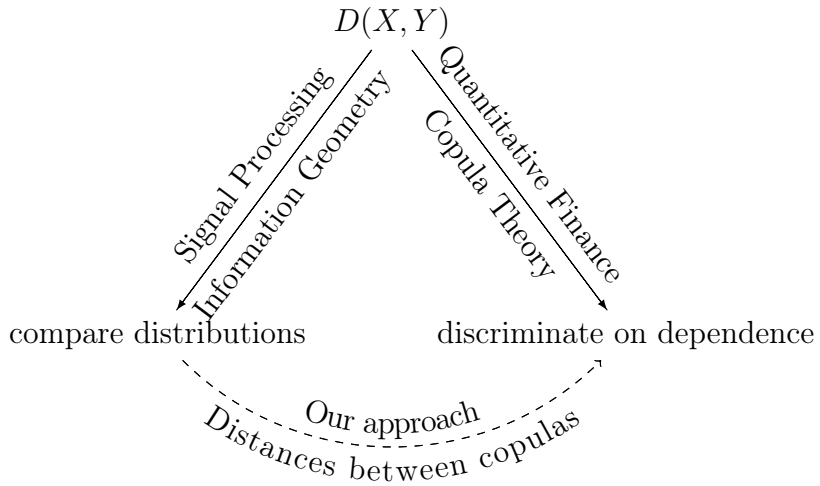


Figure 5.10: Statistical distances from Information Geometry designed to compare distributions can help for clustering time series based on their dependence. Which ones are relevant for this task? And which ones are not?

Since computing geodesics which requires solving ordinary differential equations (ODEs) can be intractable, one often considers related divergences such as Kullback-Leibler, symmetrized Jeffreys, Hellinger, or Bhattacharrya divergences which coincide with the quadratic form approximations of the Fisher-Rao distance between two close distributions, and which are computationally more tractable. These divergences all belong to the class of Ali-Silvey-Csiszár f-divergences, enjoy the information monotonicity [3] (coarsing bins decrease the divergence value), are *invariant under reparametrizations*, and furthermore induce the $\pm\alpha$-geometry for $\alpha = 3 + 2f'''(1)$ (where f is a convex function).

Alternatively to the Fisher-Rao geometry, Wasserstein distances [223] provide another natural way to compare probability distributions. Given a metric space M, these distances optimally transport the probability measure μ defined on M to turn it into ν:

$$W_p(\mu,\nu) := \left(\inf_{\gamma\in\Gamma(\mu,\nu)} \int_{M\times M} d(x,y)^p \,\mathrm{d}\gamma(x,y)\right)^{1/p},$$

where $p \geq 1$, $\Gamma(\mu,\nu)$ denotes the collection of all measures on $M\times M$ with marginals μ and ν. It can be observed that *computing the Wasserstein distance between two probability measures amounts to finding the most correlated copula associated with these measures.* Notice also that unlike Fisher-Rao and related divergences, Wasserstein distances work with probability measures instead of probability density functions.

Distances between Gaussian copulas

We illustrate the behaviour of these distances in the simple case where the underlying copula is a Gaussian (which may not be relevant for all applications). Moreover, when the compared distributions are multivariate Gaussians, we have analytical formulas (which are reported in Table 5.3).

The Gaussian copula is a distribution over the unit cube $[0,1]^d$. It is constructed from a multivariate normal distribution over $\mathbf{R}^d$ by using the probability integral transform. For a given correlation matrix $R \in \mathbf{R}^{d\times d}$, the Gaussian copula with parameter matrix R can be written as

$$C_R^{\text{Gauss}}(u_1,\ldots,u_d) = \Phi_R(\Phi^{-1}(u_1),\ldots,\Phi^{-1}(u_d)),$$

where Φ^{-1} is the inverse cumulative distribution function of a standard normal and Φ_R is the joint cumulative distribution function of a multivariate normal distribution with mean vector zero and covariance matrix equal to the correlation matrix R. For illustration purposes, we

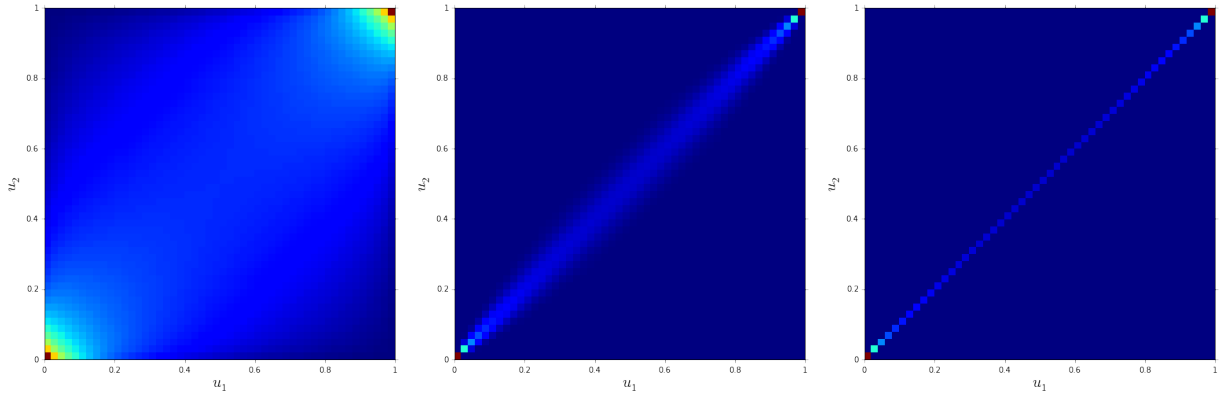


Figure 5.11: Densities of $C^{\text{Gauss}}_{R_A}, C^{\text{Gauss}}_{R_B}, C^{\text{Gauss}}_{R_C}$ respectively; Notice that for strong correlations, the density tends to be distributed very close to the diagonal.

Table 5.3: Distances in closed-form between Gaussians and their sensitivity to the correlation strength

	$D\left(\mathcal{N}(0,\Sigma_1),\mathcal{N}(0,\Sigma_2)\right)$	$D(R_A,R_B)$		$D(R_B,R_C)$
Fisher-Rao [6]	$\sqrt{\frac{1}{2}\sum_{i=1}^{n}(\log\lambda_i)^2}$	2.77	$<$	3.26
$KL(\Sigma_1\|\|\Sigma_2)$	$\frac{1}{2}\left(\log\frac{\lvert\Sigma_2\rvert}{\lvert\Sigma_1\rvert} - n + tr(\Sigma_2^{-1}\Sigma_1)\right)$	22.6	$<$	47.2
Jeffreys	$KL(\Sigma_1\|\|\Sigma_2) + KL(\Sigma_2\|\|\Sigma_1)$	24	$<$	49
Hellinger	$\sqrt{1 - \frac{\lvert\Sigma_1\rvert^{1/4}\lvert\Sigma_2\rvert^{1/4}}{\lvert\Sigma\rvert^{1/2}}}$	0.48	$<$	0.56
Bhattacharyya	$\frac{1}{2}\log\frac{\lvert\Sigma\rvert}{\sqrt{\lvert\Sigma_1\rvert\lvert\Sigma_2\rvert}}$	0.65	$<$	0.81
W_2 [202]	$\sqrt{tr\left(\Sigma_1+\Sigma_2-2\sqrt{\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}}\right)}$	**0.63**	$>$	**0.09**

λ_i eigenvalues of $\Sigma_1^{-1}\Sigma_2$; $\Sigma = \frac{\Sigma_1+\Sigma_2}{2}$

consider three bivariate Gaussian copulas parameterized by

$$R_A = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}, R_B = \begin{pmatrix} 1 & 0.99 \\ 0.99 & 1 \end{pmatrix},$$

and $R_C = \begin{pmatrix} 1 & 0.9999 \\ 0.9999 & 1 \end{pmatrix}$ respectively. Heatmaps of their densities are plotted in Fig. 5.11.

In Table 5.3, we report the distances $D(R_A,R_B)$ between $C^{\text{Gauss}}_{R_A}$ and $C^{\text{Gauss}}_{R_B}$, and the distances $D(R_B,R_C)$ between $C^{\text{Gauss}}_{R_B}$ and $C^{\text{Gauss}}_{R_C}$. We can observe that unlike Wasserstein W_2 distance, Fisher-Rao and related divergences consider that $C^{\text{Gauss}}_{R_A}$ and $C^{\text{Gauss}}_{R_B}$ are nearer than $C^{\text{Gauss}}_{R_B}$ and $C^{\text{Gauss}}_{R_C}$. This may sound surprising since $C^{\text{Gauss}}_{R_B}$ and $C^{\text{Gauss}}_{R_C}$ both describe a strong positive dependence between the two variates whereas $C^{\text{Gauss}}_{R_A}$ describes only a mild positive dependence.

Our geometric intuition to explain this fact is that Fisher-Rao geodesic distance and its related divergences are only defined on the manifold of probability distribution densities. However, the copula characterizing comonotonicity (perfect positive dependence), known as the Fréchet-Hoeffding upper bound copula $M(u_1,\ldots,u_d) = \min\{u_1,\ldots,u_d\}$, has no density. So, perfect positive dependence (for a bivariate Gaussian, it means that the two variates are perfectly correlated: $\rho = 1$) is not a point of the manifold. Unlike these distances, Wasserstein distances are defined between probability measures, so no such problem arises for the Fréchet-Hoeffding upper bound copula. In the Gaussian case considered, the closed-form formulas for these distances can make this intuition clearer. For Fisher-Rao and related

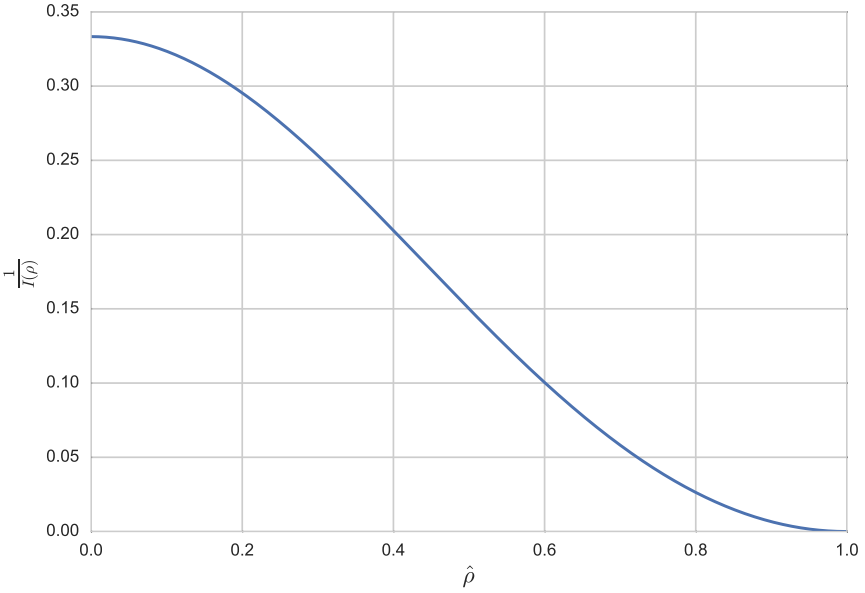


Figure 5.12: Cramér-Rao Lower Bound $\mathrm{Var}(\hat{\rho}) \geq \frac{1}{\mathcal{I}(\rho)}$ for Pearson correlation estimator

divergences, distances are expressed using the inverse of the covariance matrix and the inverse of its determinant. These matrices are ill-conditioned when correlation is strong, and singular when correlation is perfect. For Wasserstein W_2 distance, the formula is well defined in terms of square roots.

In [11], Barbaresco gives an extensive comparison of Fisher-Rao geometry versus Wasserstein geometry on the space of covariance matrices. One of the noticeable difference is that the Fisher-Rao geometry has negative curvature whereas Wasserstein geometry is flat and has nonnegative curvature. The notion of curvature is key to understand the behaviour of clustering using statistical distances. For instance, we have displayed in Fig. 5.13 the distances $D(\rho_1, \rho_2)$ between the Gaussian copulas parameterized by

$$\begin{pmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{pmatrix} \text{ and } \begin{pmatrix} 1 & \rho_2 \\ \rho_2 & 1 \end{pmatrix}.$$

One can notice that Wasserstein W_2 exhibits a roughly linear increase away from the diagonal with low curvature: It can discriminate equally well for all parameters $(\rho_1, \rho_2) \in [0, 1]^2$. On the contrary, the behaviour of Fisher-Rao strongly depends on the values ρ_1, ρ_2 as shown in Fig. 5.13: For high correlations, a small change induces a big change on the distance value due to the curvature. We call it sensitivity. In addition to returning counter-intuitive distance values as reported in Table 5.3, this property could lead to totally spurious distances and thus clusters when working with finite sample data: What if the parameters estimation error is bigger than the sensitivity? In practice, the distance would be useless.

However, Fisher-Rao and related divergences do not suffer from this drawback. They all can be locally written as a quadratic form of the Fisher Information Matrix $\mathcal{I}(\rho)$. Through this connection to the Cramér-Rao Lower Bound $\mathrm{Var}(\hat{\rho}) \geq \frac{1}{\mathcal{I}(\rho)}$ [176], they deviate (the distance sensitivity) just the right amount with respect to the statistical uncertainty of the estimator. For Pearson correlation estimate, we have $\mathrm{Var}(\hat{\rho}) \geq \frac{1}{\mathcal{I}(\rho)} = \frac{(\rho-1)^2(\rho+1)^2}{3(\rho^2+1)}$ (graphed in Figure 5.12), i.e. stronger the correlation, finer the estimate can be [144].

Proof. We consider the set of 2×2 correlation matrices $C = \begin{pmatrix} 1 & \theta \\ \theta & 1 \end{pmatrix}$ parameterized by θ.

Let $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbf{R}^2$.

$f(x;\theta) = \frac{1}{2\pi\sqrt{1-\theta^2}} \exp\left(-\frac{1}{2}x^\top C^{-1}x\right) = \frac{1}{2\pi\sqrt{1-\theta^2}} \exp\left(-\frac{1}{2(1-\theta^2)}(x_1^2 + x_2^2 - 2\theta x_1 x_2)\right)$

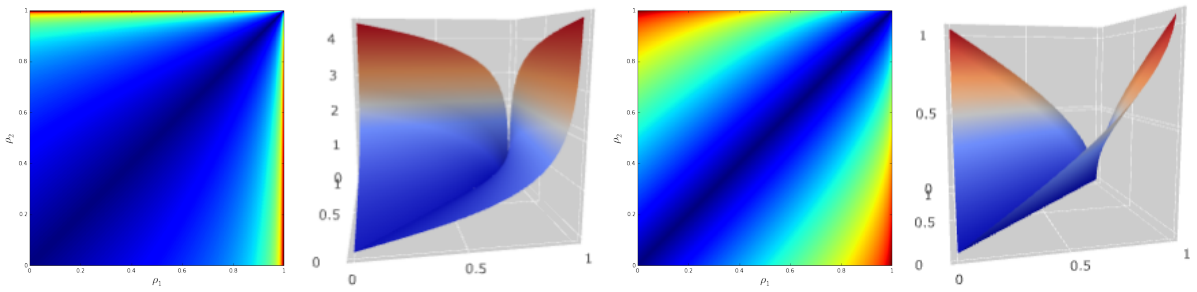


Figure 5.13: Distance heatmap and surface as a function of (ρ_1, ρ_2) for Fisher-Rao (left), for Wasserstein W_2 (right)

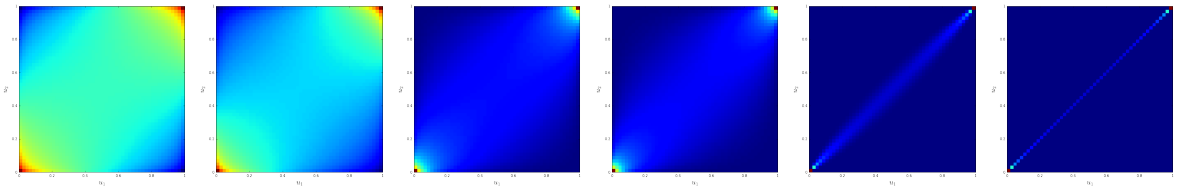


Figure 5.14: Datasets of bivariate time series are generated from six Gaussian copulas with correlation .1, .2, .6, .7, .99, .9999

$\log f(x;\theta) = -\log(2\pi\sqrt{1-\theta^2}) - \frac{1}{2(1-\theta^2)}(x_1^2 + x_2^2 - 2\theta x_1 x_2)$

$\frac{\partial^2 \log f(x;\theta)}{\partial \theta^2} = -\frac{\theta^2+1}{(\theta^2-1)^2} - \frac{x_1^2}{2(\theta+1)^3} + \frac{x_1^2}{2(\theta-1)^3} - \frac{x_2^2}{2(\theta+1)^3} + \frac{x_2^2}{2(\theta-1)^3} - \frac{x_1 x_2}{(\theta+1)^3} - \frac{x_1 x_2}{(\theta-1)^3}$

Then, we compute $\int_{-\infty}^{\infty} \frac{\partial^2 \log f(x;\theta)}{\partial \theta^2} f(x;\theta) dx$. Since $\mathbf{E}[x_1] = \mathbf{E}[x_2] = 0$, $\mathbf{E}[x_1 x_2] = \theta$, $\mathbf{E}[x_1^2] = \mathbf{E}[x_2^2] = 1$, we get

$\int_{-\infty}^{\infty} \frac{\partial^2 \log f(x;\theta)}{\partial \theta^2} f(x;\theta) dx = -\frac{\theta^2+1}{(\theta^2-1)^2} - \frac{1}{2(\theta+1)^3} + \frac{1}{2(\theta-1)^3} - \frac{1}{2(\theta+1)^3} + \frac{1}{2(\theta-1)^3} - \frac{\theta}{(\theta+1)^3} - \frac{\theta}{(\theta-1)^3} = -\frac{3(\theta^2+1)}{(\theta-1)^2(\theta+1)^2}$

Thus,

$$g(\theta) = \frac{3(\theta^2+1)}{(\theta-1)^2(\theta+1)^2}.$$

□

Clustering experiments

Fisher-Rao geodesic distance has been successfully applied for clustering and classification [76], statistical analysis (e.g., mean, median, PCA) on covariance manifolds in computational anatomy [168] and radar processing [11]. In financial applications, variates tend to be strongly correlated (for instance, correlation between maturities in a term structure can be up to 0.99). In such cases, the sensitivity problem discussed above may impair the clustering results. We illustrate this assertion by considering a dataset of N bivariate time series evenly generated from the six Gaussian copulas depicted in Fig. 5.14. When a clustering algorithm such as Ward is given a distance matrix computed from Fisher-Rao (displayed in Fig. 5.15), it will tend to gather in a cluster all copulas but the ones describing high dependence which are isolated. W_2 yields a more balanced and intuitive clustering where clusters contain copulas of similar dependence. Code for the numerical and clustering experiments are available at `www.datagrapple.com/Tech`.

Discussion

In this paper, we have focused on Gaussian copulas for two reasons: (i) we know closed-form formulas for the distances between multivariate Gaussian distributions; (ii) the existing

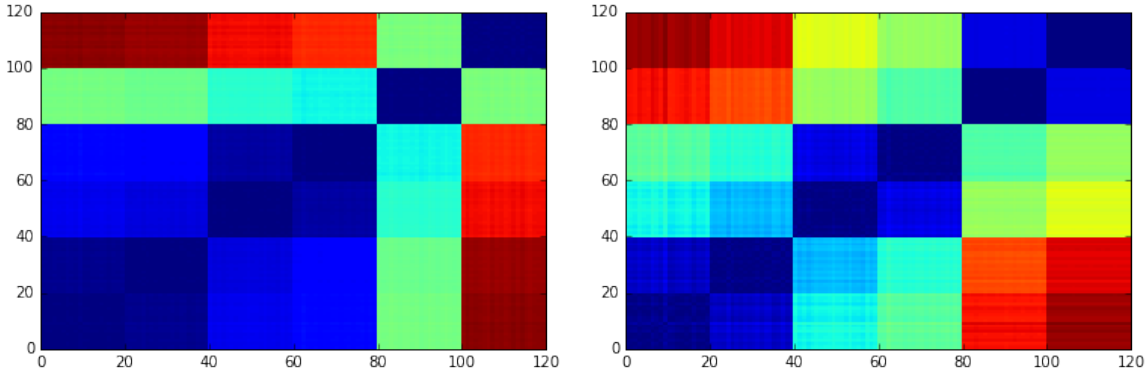


Figure 5.15: Distance heatmaps for Fisher-Rao (left), W_2 (right); Using Ward clustering, Fisher-Rao yields clusters of copulas with correlations $\{.1, .2, .6, .7\}$, $\{.99\}$, $\{.9999\}$, W_2 yields $\{.1, .2\}$, $\{.6, .7\}$, $\{.99, .9999\}$

machine learning literature focus on the manifold of covariances [1]. We have shown that if the dependence is strong between the time series, the use of Fisher-Rao geodesic distance and related divergences may not be appropriate. They are relevant to find which samples were generated from the same set of parameters (clustering viewed as a generalization of the three-sample problem [183]) due to their local expression as a quadratic form of the Fisher Information Matrix determining the Cramér-Rao Lower Bound on the variance of estimators. To measure distance between copulas, we think that the Wasserstein geometry is more appropriate since it does not lead to these counter-intuitive clusters. Beyond the Gaussian case, the phenomenon illustrated here should subsist as Fisher-Rao is defined on manifold of densities but the copula for comonotonicity cannot be part of it. We will investigate further this issue. We would also like to encompass the embedding of probability distributions into reproducing kernel Hilbert spaces [198] in our comparison of the possible distances for copulas.

Open problem: Fisher-Rao Riemannian Geometry of Correlation Matrices

Motivation:

For detecting correlation regime changes as soon as possible, Fisher-Rao geometry may be then more appropriate since its discriminative power (the manifold curvature) is function of the ease of statistical estimation depending on the values of the parameters. We may want to build a moving average of correlation matrices, compute an information ball of the last m matrices and monitor its radius: a relatively large radius implies significant changes, a relatively small radius implies no significant changes. The mathematical difficulty is to compute an appropriate Riemannian mean of the m correlation matrices.

Formulation of the problem:

The goal is to provide the Fisher-Rao Riemannian metric, geodesic, distance and mean for the set of correlation matrices. Such results will allow *intrisic* computing in the set of correlation matrices, and thus may improve the methodology of many applications in signal processing, e.g. radar, telecom, finance.

Introduction

Considering the set of centered normal distributions $\{\mathcal{N}(0, C) \mid C \in \mathcal{E}\}$, where C lives in the elliptope, i.e. the set of correlation matrices:

$$\mathcal{E} = \{C \in \mathbf{R}^{n \times n} \mid C = C^\top, \forall x \in \mathbf{R}^n, x^\top C x \geq 0, \forall i \in \{1, \ldots, n\}, C_{ii} = 1\}$$

Example. For $n = 3$, correlation matrices can be written $\begin{pmatrix} 1 & x & y \\ x & 1 & z \\ y & z & 1 \end{pmatrix}$ verifying above conditions. $\mathcal{E}$ is the compact convex set whose boundary is displayed in Figure 5.16. Correlation matrices of rank 1 are the corners, of rank 2 the faces, and rank 3 are inside. Our goal is to endow $\mathcal{E}$ with a metric having good properties. For example, we may want that the mean of two correlation matrices of rank 2 is a correlation matrix of rank 2. More generally, we may want that a geodesic, i.e. a 'shortest path', between two such matrices stays in this subset. This is not the case with Euclidean geometry.

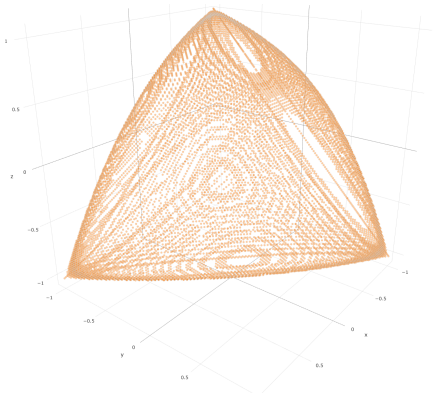


Figure 5.16: 3D elliptope. Each 3×3 correlation matrix is represented by a point (x, y, z)

Considering the set of centered normal distributions $\{\mathcal{N}(0, \Sigma) \mid \Sigma \in \mathcal{C}\}$, where Σ lives in the cone of PSD matrices, i.e. the set of covariance matrices:

$$\mathcal{C} = \{\Sigma \in \mathbf{R}^{n \times n} \mid \Sigma = \Sigma^\top, \forall x \in \mathbf{R}^n, x^\top \Sigma x \geq 0\}$$

$\mathcal{E} \subset \mathcal{C}$.

Example. In Figure 5.17, we display the boundary of the set of 2×2 covariances matrices $\begin{pmatrix} x & y \\ y & z \end{pmatrix}$. The blue segment is the subset $\mathcal{E}$.

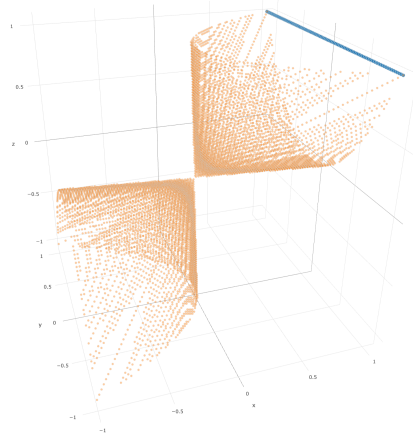


Figure 5.17: 3D cone. Each 2×2 covariance matrix is represented by a point (x, y, z). The blue segment $(x = z = 1)$ is the set of 2×2 correlation matrices

Fisher-Rao Riemannian Geometry of Covariance Matrices

Information Geometry for multivariate Gaussian with zero mean and intrinsic geometry of positive-semidefinite matrices lead to same metric and distance.

- metric: $ds^2 = \|\Sigma^{-1/2} d\Sigma \Sigma^{-1/2}\|^2 = \mathrm{Tr}\left((\Sigma^{-1} d\Sigma)^2\right)$
- distance: $d^2(\Sigma_1, \Sigma_2) = \frac{1}{2}\sum_{i=1}^{n} \log(\lambda_i)^2$, λ_i eigenvalues of $\Sigma_1^{-1}\Sigma_2$
- geodesic between Σ_1 and Σ_2, $t \in [0,1]$: $\gamma(t) = \Sigma_1^{1/2} e^{t \log\left(\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2}\right)} \Sigma_1^{1/2}$
- mean of Σ_1 and Σ_2: $\gamma(1/2)$
- mean of $\Sigma_1, \ldots, \Sigma_n$: $\mathrm{argmin}_{\Sigma \in \mathcal{C}} \sum_{k=1}^{n} d^2(\Sigma, \Sigma_i)$

These results are well known and widely used in signal processing. Some use them for correlation matrices (since they are also covariance matrices),

but...

The submanifold $\mathcal{E}$ is not totally geodesic in $\mathcal{C}$

We can observe that the submanifold $\mathcal{E}$ is not totally geodesic in $\mathcal{C}$ which means that a geodesic (using the metric from $\mathcal{C}$) between two points in $\mathcal{E}$ do not necessarily live in $\mathcal{E}$ (cf. Figure 5.18). As a consequence, though we can use the geometry for covariance matrices, the mean of correlation matrices is not necessarily a correlation matrix but a covariance matrix.

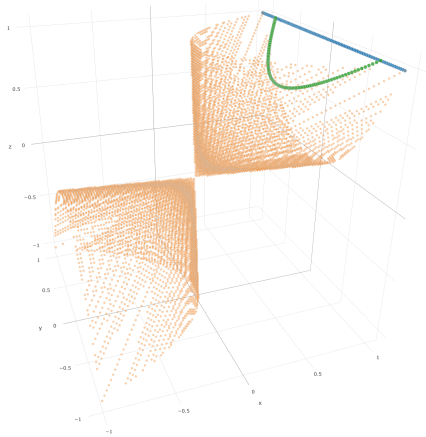


Figure 5.18: 3D cone. Each 2×2 covariance matrix is represented by a point (x, y, z) The blue segment ($x = z = 1$) is the set of 2×2 correlation matrices. In green, the geodesic (using Fisher-Rao metric for covariances) between correlation matrix $(1, -0.75, 1)$ and correlation matrix $(1, 0.75, 1)$. It is not included in the blue segment representing $\mathcal{E}$

Open questions

Find for $\mathcal{E}$, how to compute

- its metric?
- associated distance?
- its geodesics?
- a mean?

Optimal Transport

In [144], authors illustrate in a parametric setting using Gaussian copulas that common divergences (such as Kullback-Leibler, Jeffreys, Hellinger, Bhattacharyya) are not relevant for clustering these distributions, especially when dependence is high. These information divergences are only defined for absolutely continuous measures whereas some copulas have no density (e.g. the one for positive dependence). In practice, when working with frequency histograms, it gets worse: One has to pre-process the empirical measures with a kernel density estimator before computing these divergences. On the contrary, optimal transport distances are well-defined for both discrete (e.g. empirical) and continuous measures.

The idea of optimal transport is intuitive. It was first formulated by Gaspard Monge in 1781 [150] as a problem to efficiently level the ground: Given that work is measured by the distance multiplied by the amount of dirt displaced, what is the minimum amount of work required to level the ground? Optimal transport plans and distances give the answer to this problem.

In practice, empirical distributions can be represented by histograms. We follow notations from [59]. Let r, c be two histograms in the probability simplex $\Sigma_m = \{x \in \mathbb{R}_+^m : x^\top 1_m = 1\}$. Let $U(r,c) = \{P \in \mathbb{R}_+^{m\times m} \mid P1_m = r, P^\top 1_m = c\}$ be the transportation polytope of r and c, that is the set containing all possible transport plans between r and c.

Definition 12 (Optimal Transport)**.** *Given a $m \times m$ cost matrix M, the cost of mapping r to c using a transportation matrix P can be quantified as $\langle P, M\rangle_F$, where $\langle \cdot, \cdot \rangle_F$ is the Frobenius dot-product. The optimal transport between r and c given transportation cost M is thus:*

$$d_M(r,c) := \min_{P \in U(r,c)} \langle P, M\rangle_F. \tag{5.16}$$

Whenever M belongs to the cone of distance matrices, the optimum of the transportation problem $d_M(r,c)$ is itself a distance.

Lightspeed transportation. Optimal transport distances suffer from a computational burden scaling in $O(m^3 \log m)$ which has prevented their widespread use in machine learning: A mere distance computation between two high-dimensional histograms can take several seconds. In [59], Cuturi provides a solution to this problem: He restrains the polytope $U(r,c)$ of all possible transport plans between r and c to a Kullback-Leibler ball $U_\alpha(r,c) \subset U(r,c)$, where $U_\alpha(r,c) = \{P \in U(r,c) \mid \mathrm{KL}(P\|rc^\top) \leq \alpha\}$. He then shows that it amounts to perform an entropic regularization of the optimal transportation problem whose solution is smoother and less deterministic. The regularized optimal transportation problem is now strictly convex, and can be solved efficiently using the Sinkhorn-Knopp iterative algorithm which exhibits linear convergence. Its solution is the Sinkhorn distance [59]:

$$d_{M,\alpha}(r,c) := \min_{P \in U_\alpha(r,c)} \langle P, M\rangle_F, \tag{5.17}$$

and its dual $d_M^\lambda(r,c)$: $\forall \alpha > 0, \exists \lambda > 0$,

$$d_{M,\alpha}(r,c) = d_M^\lambda(r,c) := \langle P^\lambda, M\rangle_F, \tag{5.18}$$

where $P^\lambda = \mathrm{argmin}_{P\in U(r,c)} \langle P, M\rangle_F - \frac{1}{\lambda} h(P)$, and h is the entropy function.

In the following, we will leverage the dual-Sinkhorn distances for comparing, clustering and computing the clusters centers [60] of a set of copulas at full speed.

A methodology to explore and measure non-linear correlations

We propose an approach to explore and measure non-linear correlations between N variables $X_1, \ldots, X_N$ in a dataset. These N variables can be, for instance, time series or features. The methodology presented (which is summarized in Figure 5.19) is twofold, and consists of: (i) an exploratory part of the pairwise dependence between variables, (ii) the parameterization and use of a novel dependence coefficient.

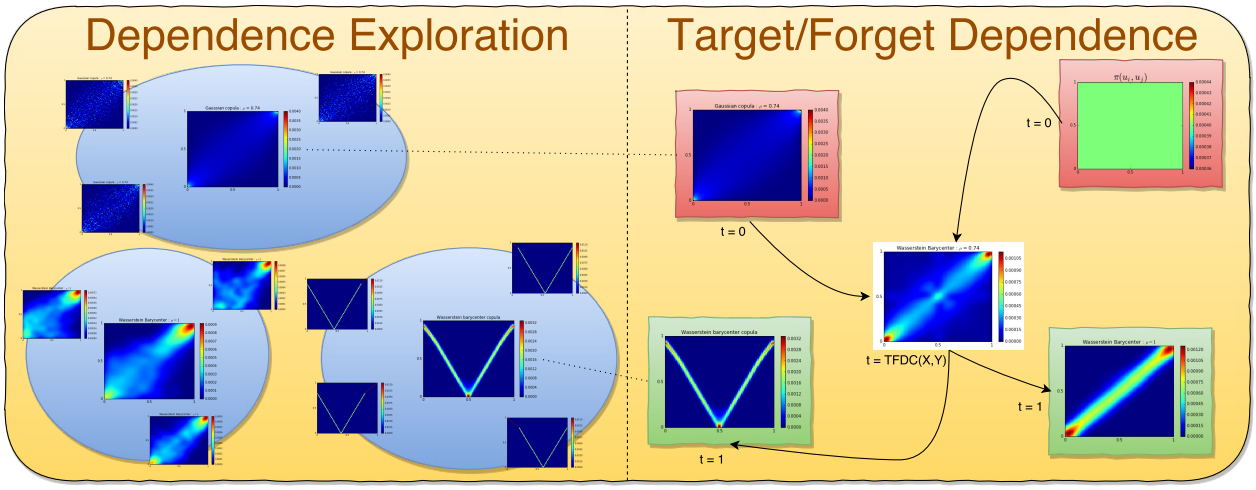


Figure 5.19: Exploration (left panel) and measure (right panel) of non-linear correlations. Exploration consists in finding clusters of similar copulas, visualizing their centroids, and eventually using them to assess the dependence of given variables represented by their copula

Using transportation of copulas as a measure of correlations

In this section, we leverage and extend the idea presented in our short introduction to copulas: correlation coefficients can be viewed as a distance between the data-copula and the Fréchet-Hoeffding bounds or the independence copula. The distance involved is usually an ℓ_p Minkowski metric distance. In the following, we will:

- replace the ℓ_p distance by an optimal transport distance between measures,
- parameterize a dependence coefficient with other copulas than the Fréchet-Hoeffding bounds or the independence one.

Using the optimal transport distance between copulas, we now propose a dependence coefficient which is parameterized by two sets of copulas: *target* copulas and *forget* copulas.

Definition 13 (Target/Forget Dependence Coefficient)**.** *Let $\{C_l^-\}_l$ be the set of forget-dependence copulas. Let $\{C_k^+\}_k$ be the set of target-dependence copulas. Let C be the copula of (X_i, X_j). Let d_M be an optimal transport distance parameterized by a ground metric M. We define the Target/Forget Dependence Coefficient as such:*
$\text{TFDC}\left(X_i, X_j; \{C_k^+\}_k, \{C_l^-\}_l\right) :=$

$$\frac{\min_l d_M(C_l^-, C)}{\min_l d_M(C_l^-, C) + \min_k d_M(C, C_k^+)} \in [0,1]. \tag{5.19}$$

Using this definition, we obtain: $\text{TFDC}\left(X_i, X_j; \{C_k^+\}_k, \{C_l^-\}_l\right) = 0 \Leftrightarrow C \in \{C_l^-\}_l$, $\text{TFDC}\left(X_i, X_j; \{C_k^+\}_k, \{C_l^-\}_l\right) = 1 \Leftrightarrow C \in \{C_k^+\}_k$.

Example. A standard correlation coefficient can be obtained by setting the forget-dependence set to the independence copula, and the target-dependence set to the Fréchet-Hoeffding bounds. How does it compare to the Spearman correlation? In Figure 5.20, we display how the two coefficients behave on a simple numerical experiment: $X = Z\mathbf{1}_{Z<a} + \epsilon_X \mathbf{1}_{Z>a}$, $Y = Z\mathbf{1}_{Z<a+0.25} + \epsilon_Y \mathbf{1}_{Z>a+0.25}$, where Z is uniform on $[0,1]$ and ϵ_X, ϵ_Y are independent noises. That is $X = Y$ over $[0, a]$. Notice that for $a = 0.75$, Spearman coefficient takes a negative value. We may thus prefer the monotonically increasing behaviour of the TFDC to the Spearman one.

How to choose, design and build targets?

We now propose two alternatives for choosing, designing and building the *target* and *forget* copulas: an exploratory data-driven approach and an hypotheses testing approach.

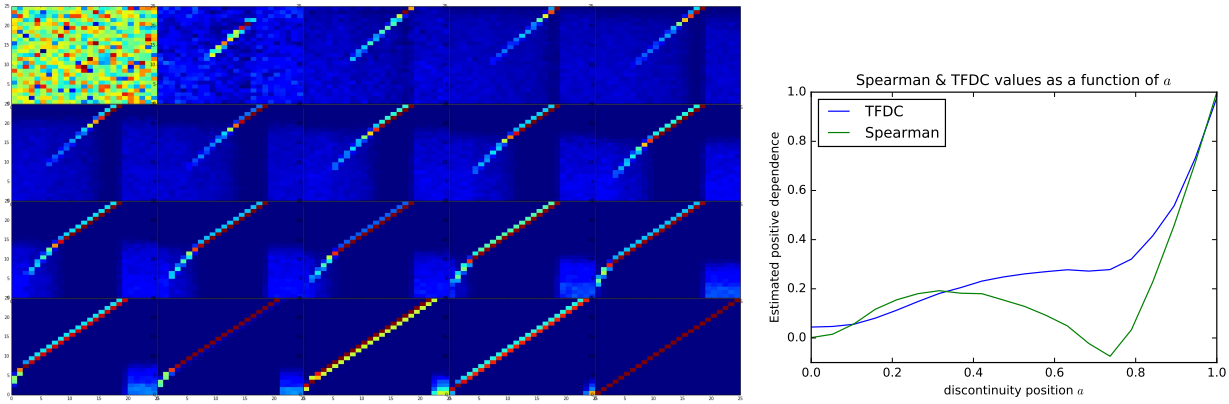


Figure 5.20: Empirical copulas for (X, Y) where $X = Z\mathbf{1}_{Z<a} + \epsilon_X \mathbf{1}_{Z>a}$, $Y = Z\mathbf{1}_{Z<a+0.25} + \epsilon_Y \mathbf{1}_{Z>a+0.25}$, $a = 0, 0.05, \ldots, 0.95, 1$, and where Z is uniform on $[0, 1]$ and ϵ_X, ϵ_Y are independent noises (left figure). Top left is an empirical copula for independence ($a = 0$), bottom right is the copula for perfect positive dependence ($a = 1$). Parameter a is increasing from top to bottom, and from left to right; TFDC and Spearman coefficients estimated between X and Y as a function of a (right figure). For $a = 0.75$, Spearman coefficient yields a negative value, yet $X = Y$ over $[0, a]$

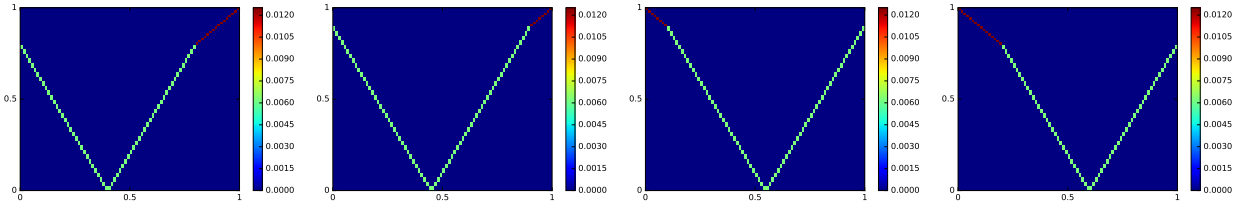


Figure 5.21: 4 copulas describing the dependence between $X \sim \mathcal{U}([0, 1])$ and $Y \sim (X \pm \epsilon_i)^2$, where ϵ_i is a constant noise specific for each distribution. X and Y are counter-monotonic (more or less) half of the time, and co-monotonic (more or less) half of the time

Data-driven: Clustering of copulas

Assume we have N variables $X_1, \ldots, X_N$, and T observations for each of them. First, we compute $\binom{N}{2} = O(N^2)$ empirical copulas which represent the dependence structure between all the couples (X_i, X_j). Then, we summarize all these distributions using a center-based clustering algorithm, and extract the clusters centers using a fast computation of Wasserstein barycenters [60]. A given center represents the mean dependence between the couples (X_i, X_j) inside the corresponding cluster. Figure 5.21 and 5.22 illustrate why a Wasserstein W_2 barycenter, i.e. the minimizer $\mu^\star$ of $\frac{1}{N}\sum_{i=1}^{N} W_2^2(\mu, \nu_i)$ [2] where $\{\nu_1, \ldots, \nu_N\}$ is a set of N measures (here, bivariate empirical copulas), is more relevant to our needs: we benefit from robustness against small deformations of the dependence patterns.

Example. In Table 5.4, we display some interesting dependence patterns which can be found in UCI datasets `http://archive.ics.uci.edu/ml/`. In this case, variables $X_1, \ldots, X_N$ are the N features. Some associations are easy to explain (e.g. top left copula representing the relation between radius and area of roughly round cells in the `Breast Cancer Wisconsin (Diagnostic) Data Set`) whereas some others less (e.g. top row third copula from the left which represents the relation between the perimeter and the fractal dimension of the cells).

An equitable copula-based dependence measure such as the one described in [88] may detect them well, but will also detect the spurious ones which are due to artifacts in the data (or pure chance). With this approach, one can spot them and add them to the set of forget-dependence copulas. For these reasons, we think that this approach could improve the feature selection correlation-based approaches [92, 228] which rely on the hypothesis that

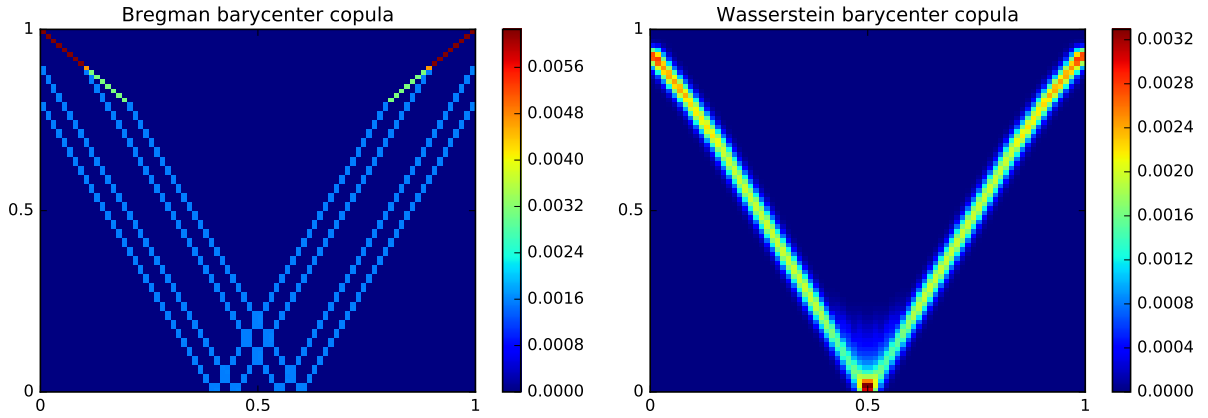
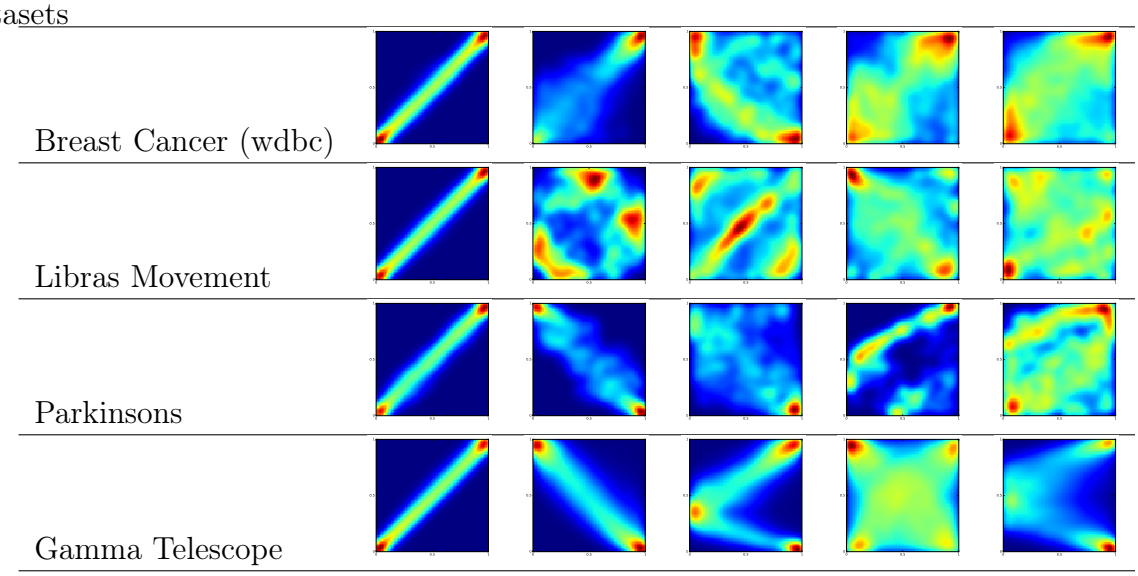


Figure 5.22: Barycenter of the 4 copulas from Figure 5.21 for: (left) Bregman geometry [10] (which includes, for example, squared Euclidean and Kullback-Leibler distances); (right) Wasserstein geometry. Notice that the Wasserstein barycenter better describes the underlying dependence between X and Y: the copula encodes a functional association. This is not the case for the Bregman barycenter

good feature subsets contain features highly correlated with the class, yet uncorrelated with each other [92].

Table 5.4: Dependence patterns (= clustering centroids) found between variables in UCI datasets



Targets as hypotheses from an expert

One can specify dependence hypotheses, generate the corresponding copulas, then measure and rank correlations with respect to them. For example, one can answer to questions such as: Which are the pairs of assets that are usually positively correlated for small variations but uncorrelated otherwise? In [73], authors present a method for constructing bivariate copulas by changing the values that a given copula assumes on some subrectangles of the unit square. They discuss some applications of their methodology including the construction of copulas with different tail dependencies. Building *target* and *forget* copulas is another one. In the Experiments section, we illustrate its use to answer the previous question and other dependence queries.

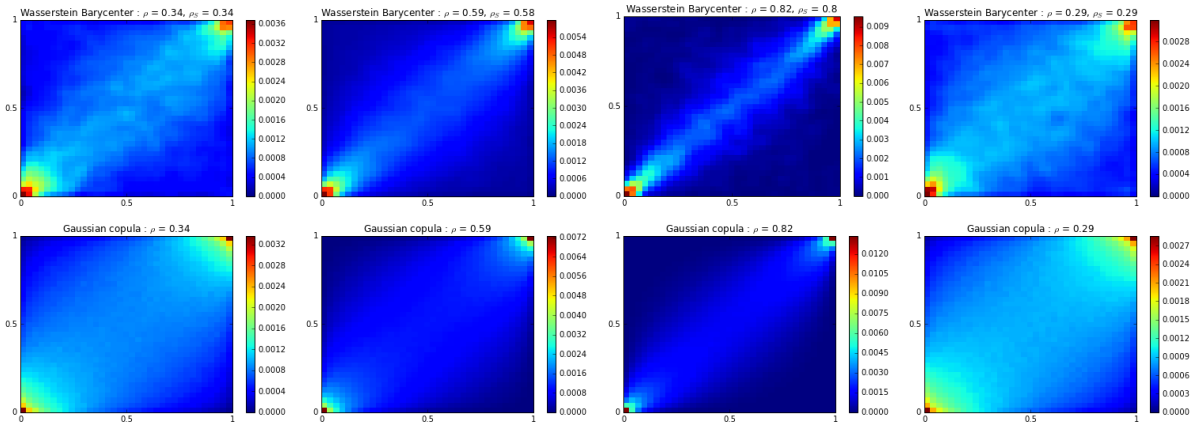
Experiments

Exploration of financial correlations

We illustrate the first part of the methodology with three different datasets of financial time series. These time series consist in the daily returns of stocks (40 stocks from the CAC 40 index comprising the French highest market capitalizations), credit default swaps (75 CDS from the iTraxx Crossover index comprising the most liquid sub-investment grade European entities) and foreign exchange rates (80 FX rates of major world currencies) between January 2006 and August 2016. We display some of the clustering centroids obtained for each asset class on the top row, and below we display their corresponding Gaussian copulas parameterized by the estimated linear correlations. Notice the strong difference between the empirical copulas and the Gaussian ones which are still widely used in financial engineering due to their convenience. Notice also the difference between asset classes: Though estimated correlations are $\rho = 0.34$ for the leftmost copulas, they have much dissimilar peculiarities.

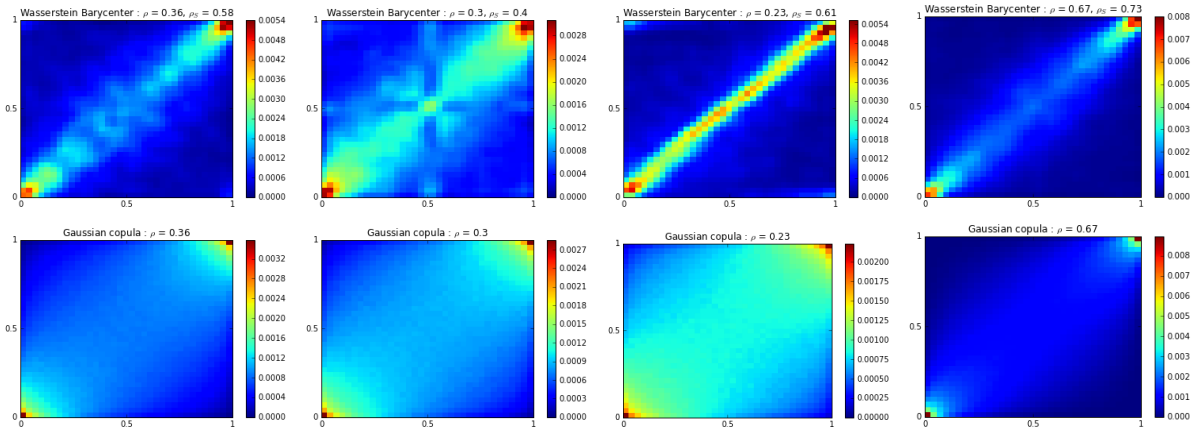
Stocks

Centroids' main feature: More mass in the bottom-left corner, i.e. lower tail dependence. Stock prices tend to plummet together.



Credit default swaps

Centroids' main feature: More mass in the top-right corner, i.e. upper tail dependence. Insurance cost against entities' default tends to soar in stressed market.



FX rates

Centroids' main feature: Empirical copulas show that dependence between FX rates are various. For example, rates may exhibit either strong dependence or independence while being anti-correlated during extreme events.

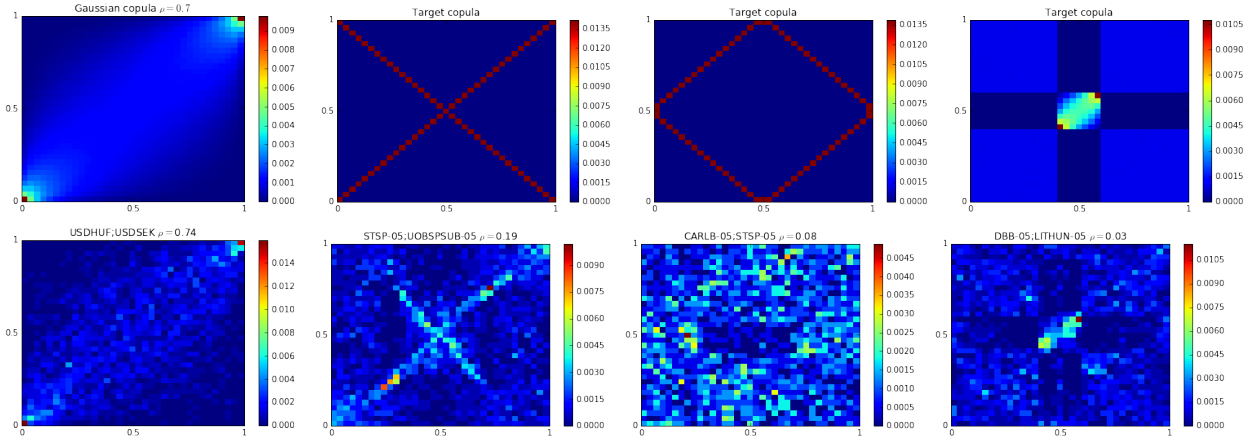
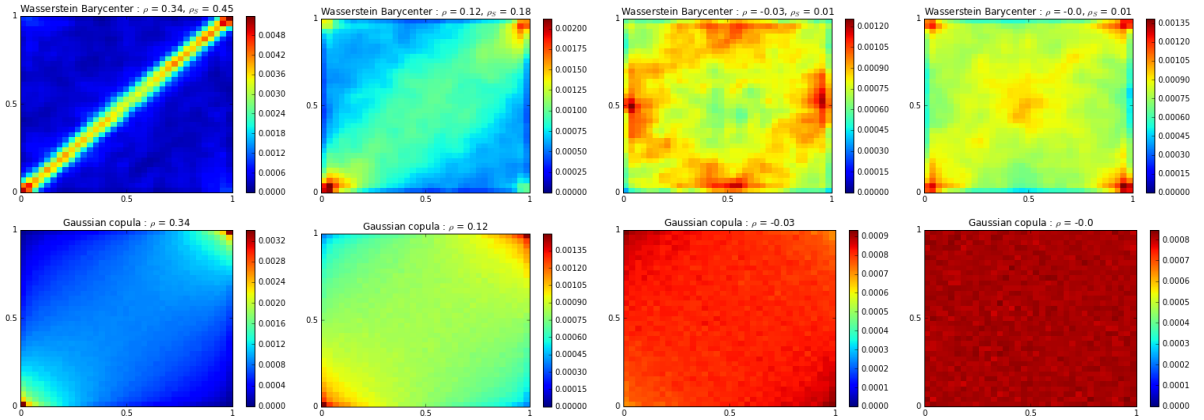


Figure 5.23: Target copulas (simulated or handcrafted) and their respective nearest copulas which answer questions A,B,C,D



Answering dependence queries

Inspired by the previous exploration results, we may want to answer such questions: (A) Which pair of assets having $\rho = 0.7$ correlation has the nearest copula to the Gaussian one? Though such questions can be answered by computing a likelihood for each pairs, our methodology stands out for dealing with non-parametric dependence patterns, and thus for questions such as: (B) Which pairs of assets are both positively and negatively correlated? (C) Which assets occur extreme variations while those of others are relatively small, and conversely? (D) Which pairs of assets are positively correlated for small variations but uncorrelated otherwise?

Considering a cross-asset dataset which comprises the SBF 120 components (index including the CAC 40 and 80 other highly capitalized French entities), the 500 most liquid CDS worldwide, and 80 FX rates, we display in Figure 5.23 the empirical copulas (alongside their respective targets) which best answer questions A,B,C,D.

Power of TFDC

In this experiment, we compare the empirical power of TFDC to well-known dependence coefficients such as Pearson linear correlation (cor), distance correlation (dCor) [201], maximal information coefficient (MIC) [179], alternating conditional expectations (ACE) [27], maximum mean discrepancy (MMD) [91], copula maximum mean discrepancy (CMMD) [88], randomized dependence coefficient (RDC) [132]. Statistical power of a binary hypothesis test is the probability that the test correctly rejects the null hypothesis (H0) when the alternative hypothesis (H1) is true. In the case of dependence coefficients, we consider (H0): X and Y are independent; (H1): X and Y are dependent. Following the numerical experiment described in [191, 132], we estimate the power of the aforementioned dependence measures

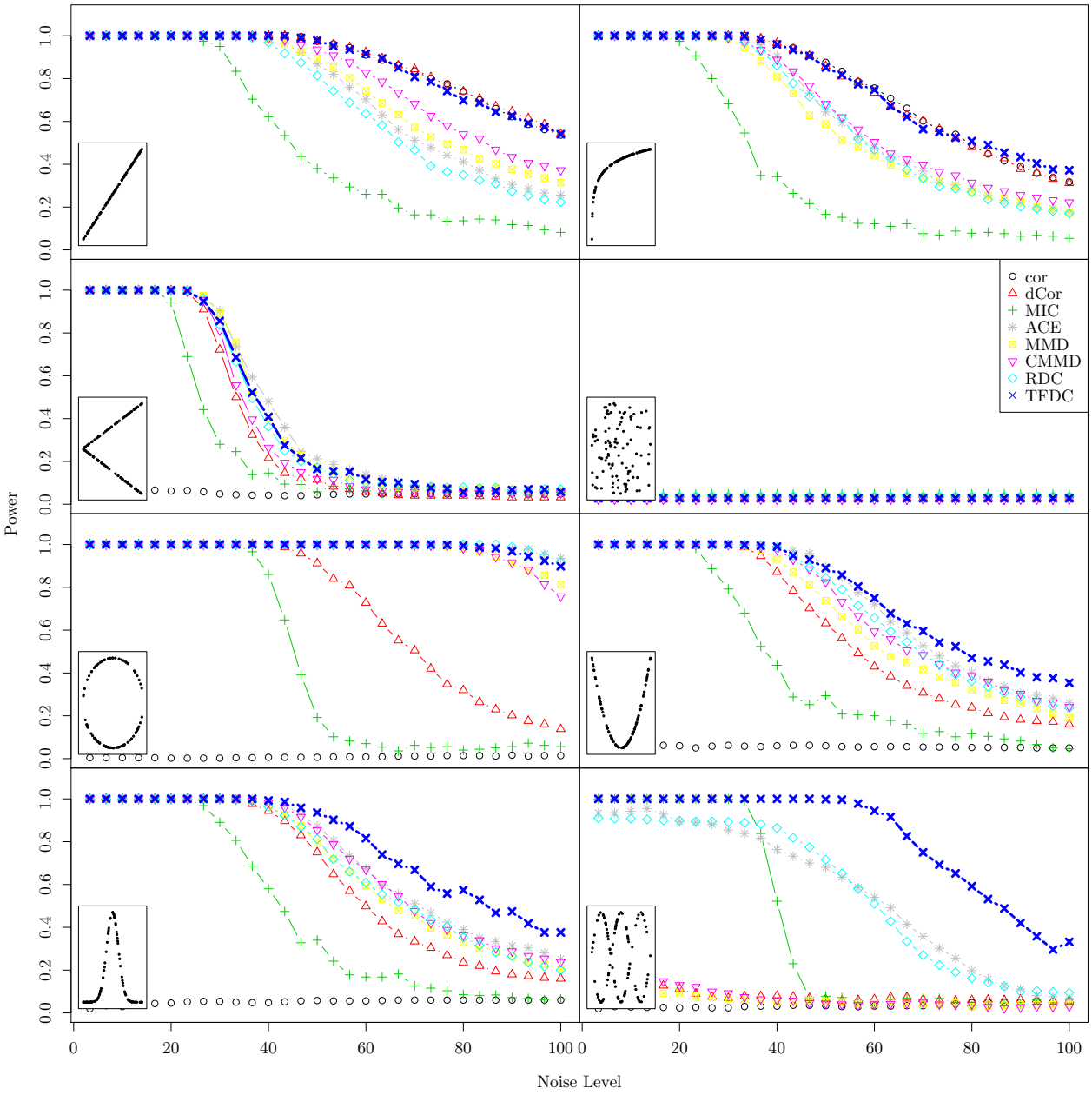


Figure 5.24: Power of several dependence coefficients as a function of the noise level in eight different scenarios. Insets show the noise-free form of each association pattern. The coefficient power was estimated via 500 simulations with sample size 500 each

with simulated pairs of variables with different relationships (considered in [179, 191, 132]), but with varying levels of noise added. By design, TFDC aims at detecting the simulated dependence relationships. Thus, this dependence measure is expected to have a much higher power than coefficients such as MIC since, according to Simon and Tibshirani in [191], coefficients "which strive to have high power against all alternatives can have low power in many important situations." TFDC only targets the specific important situations. Results are displayed in Figure 5.24.

Discussion

It is known by risk managers how dangerous it can be to rely solely on a correlation coefficient to measure dependence. That is why we have proposed a novel approach to explore, summarize and measure the pairwise correlations which exist between variables in a dataset. We have also pointed out through the UCI-datasets example that non-trivial dependence patterns can be easily found between the features variables. Using these patterns as *targets*

when performing correlation-based feature selection may improve results. This idea still needs to be empirically verified. The experiments show the benefits of the proposed method: It allows to highlight the various dependence patterns that can be found between financial time series, which strongly depart from the Gaussian copula widely used in financial engineering. Though *answering dependence queries* as briefly outlined is still an art, we plan to develop a rich language so that a user can formulate complex questions about dependence, which will be automatically translated into copulas in order to let the methodology provide these questions accurate answers.

Chapter 6

Practical considerations for using clustering

6.1 Number of clusters

"What is the correct number of clusters?" is the perennial question for practitioners. Many papers have been published on this topic [199, 116, 149, 185, 206] (and the list goes on and on), but no off-the-shelf estimator working well in practice and for a broad range of domains has been found yet.

As we have seen throughout this thesis, the similarity/correlation matrix is highly hierarchical, it is not clear that there is a unique 'correct' number of clusters but there are maybe layers in the hierarchy which are more relevant than others. We propose to use a simple bootstrapping procedure [75] to try to find them: We generate 100 bootstrapped samples of the initial time series ($N = 1025$ CDS time series of length $T = 750$ (3 years)) and compute the 100 corresponding clustering for a given number of clusters K. We associate to K a score which is the average similarity (measured using the Adjusted Rand Index (ARI)) of the obtained partitions. The more similar these partitions, the more relevant is the tested K, as the results are robust to small perturbations of the data which is a desirable property. We display in Figure 6.1 the similarity scores obtained daily during two months for the different values of K. We can notice that these curves are stable from one day to another and that they indicate some layers which are more stable than their adjacent layers (local maxima). However, we can see that the hierarchy is quite developed: unlike clusters built from datasets in other domains which become unstable for high values of K, we have noticed that the stability score was increasing for values of K tending to the number of time series N (then going to 0 for $K = N$ (like for $K = 1$) the trivial cases). Indeed, we can understand that the European automotive sub-sector is a strong meaningful cluster, but some pairs like Peugeot/Renault or BMW/Daimler are even more robust and meaningful.

Ultimately, the correct number of clusters has to be task dependent: For example, if one wants to devise a cluster-based mean reverting strategy using potential cointegration relationships, then one should probably select a fine partition giving relatively small and very correlated clusters. Using these clusters should help to avoid using spurious and fast breaking cointegration relationships.

6.2 Imputation of missing values in multivariate time series

We may want that the N time series of length T have no missing values. For example, this can be an assumption of risk models to work with full historical data (no missing values in the $N \times T$ data matrix). In the CDS market, however, it can happen that some single names are not quoted every day. It can also happen that new credit default swaps are created

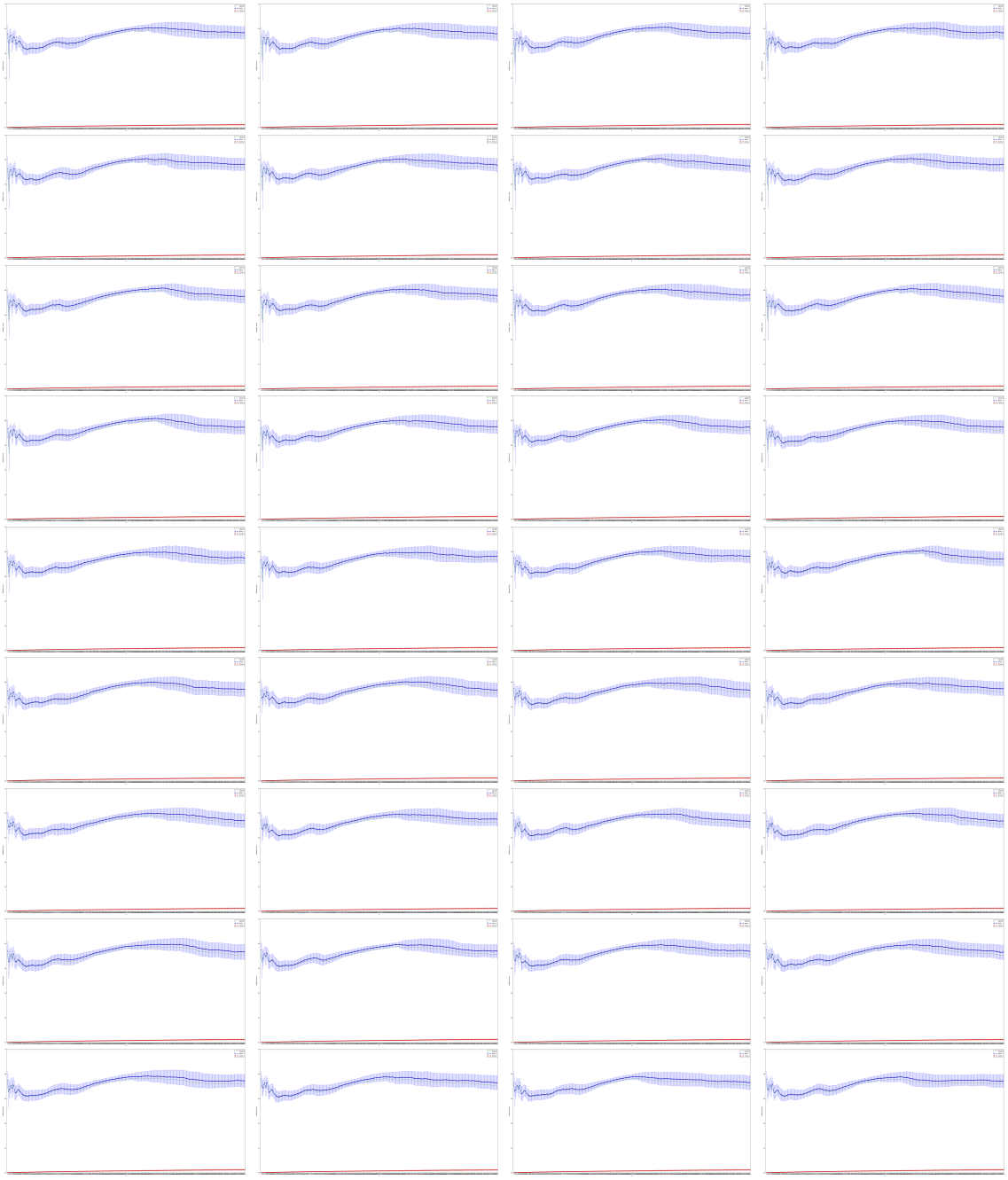


Figure 6.1: Stability scores for clustering $N = 1025$ CDS time series of length $T = 750$ (3 years) as a function of the number of clusters K, $K \in [\![2, 343]\!]$, between 13-08-2016 and 07-11-2016. These scores are stable from one day to another. Local maxima indicate layers in the hierarchy with a stronger stability. They can be found at 2 (essentially Europe vs. US), 6 (essentially the economic regions (Europe, US, Asia) subdivided by the quality of credit: investment grade vs. high-yield), 10 (US and Europe further subdivided by corporate vs. financial). The red curves correspond to the stability scores of clustering pure noise (T i.i.d. samples of a N-variate Gaussian $\mathcal{N}(0, \Sigma)$, where $\Sigma = I_N$). Clusters of pure noise are not stable (we can notice however a small bias of the stability score for large values of K).

and therefore have no historical data. How can we seamlessly incorporate them in existing risk/trading/analysis frameworks which assume the complete history hypothesis?

Cluster-based regressions

In this section, we propose to use clustering as a relevant way to impute the missing values for these time series. We suggest using the following procedure:

- for time series having enough observations (let's say 120 quoted days), we use **statistical benchmarks**, i.e. clusters obtained by one of the many clustering methodologies, to impute the missing prices. We first build statistical clusters based on time series having full history, then we find the nearest cluster for each of the missing-values time series based on their history, and finally we do a regression between the existing quotes and the center of their respective cluster;
- for time series having too few observations (let's say less than 120), we use **economic benchmarks**, i.e. clusters crafted by macro and economic variables, to build the regression.

Optional step specific to credit default swaps: A quadratic process for time series of prices normalization

As conventional spreads may not be compatible with any upfront price, we prefer to use directly upfront quotations. As the regression aims at completing the missing values of the time series, it is helpful to have a statistical model that describes the upfront-prices behaviour correctly. We use a quadratic diffusion model that proved to be efficient for option pricing and Value-at-Risk computations.

At any quotation date t, the clean upfront price $NPV(t)$ can be written:

$$NPV(t) = (S(t) - C)Rbp(t),$$

where $Rbp(t)$ is the risky duration inferred by a probabilistic model, C is the quoted coupon, and $S(t)$ is the par-spread. If $R(t)$ is the residual maturity,

$$NPV(t) + CR(t) = S(t)Rbp(t) + C(R(t) - Rbp(t)) \in [0, 1 + CR(t)]$$

as $S(t)Rbp(t)$ is the market expected loss and $R(t) \geq Rbp(t)$.

We consider the following quantity:

$$U(t) = \frac{NPV(t) + CR(t)}{1 + CR(t)} \in [0, 1].$$

The quadratic diffusion process allows to model time series belonging to $[0, 1]$:

$$dU(t) = U(t)(1 - U(t))\sigma dW(t).$$

We therefore choose to work on the following series of innovations $N(t)$:

$$N(t) = \frac{U(t) - U(t-1)}{U(t)(1 - U(t))}.$$

Results

As time goes on this procedure is applied regularly to refine the imputed missing values. For new entities, past values are imputed first using economic clusters (cf. Figure 6.2 and blog post `https://www.datagrapple.com/Blog/Show/11831/spring-rolls.html`), then after 6 months they can be estimated using statistical clusters (cf. Figures 6.3, 6.4). With more data, it is possible that the cluster assignment changes and therefore the estimated past.

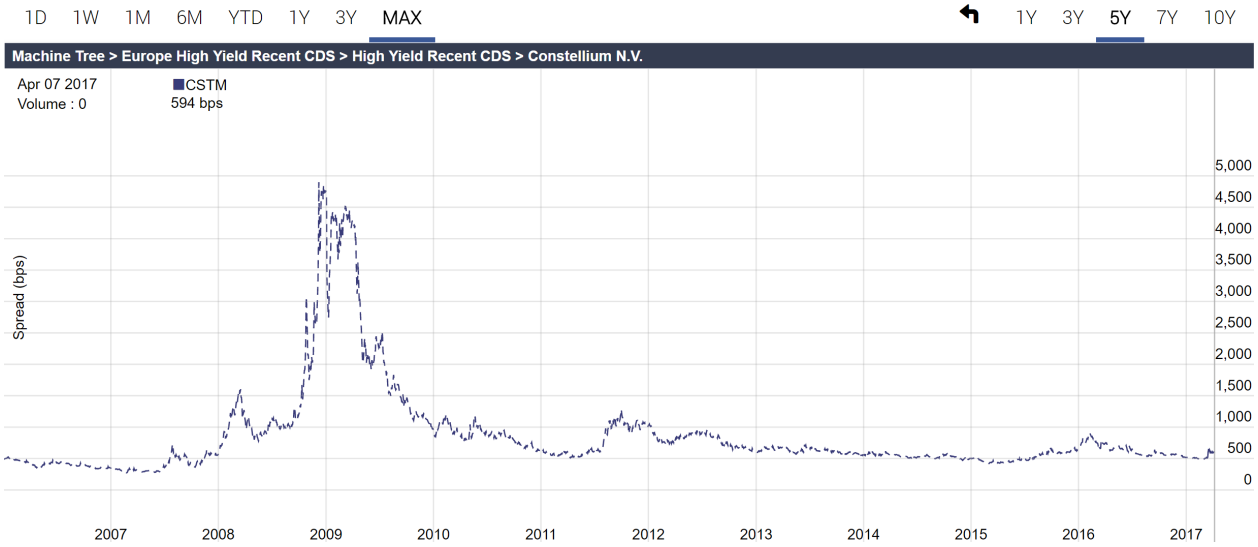


Figure 6.2: CSTM (Constellium NV) - the producer of aluminium cans but also elements for planes and cars (CSTM includes former French Pechiney activities and Arconic is its main competitor) - has been added to the iTraxx Crossover index and thus should start to be traded actively. This entity has no previous CDS history.

Figure 6.3: This statistical cluster gathers high-yield European entities which have only partial historical data. Bounded by the red box, we can see in dotted lines the spreads imputed by the cluster-based regression for MTNLN (Matalan Finance PLC), CAREUK (Care UK Health and Social Care PLC), SELNSW (Selecta Group B.V.), PFDLN (Premier Foods Finance PLC), HEMABV (Hema Bondco I B.V.), PIZEXP (PizzaExpress Financing 1 PLC), ICELTD (Iceland Bondco PLC), BROPRLN (Boparan Finance PLC), CABBCO (Monitchem Holdco 3 S.A.), DRYMIX (Dry Mix Solutions Investissements S.A.S.), STGATE (Stonegate Pub Company Financing PLC), CVRD (VALE S.A.), GALAPG (Galapagos Holding S.A.), LINDOR (Lock Lower Holding AS), NUMFP (Numericable-SFR S.A.), NVFVES (Novafives S.A.S) which were introduced in the iTraxx Crossover index during the September 2014 roll, and ZIGGO (Ziggo Bond Finance B.V.) which was introduced during the next roll in March 2015.

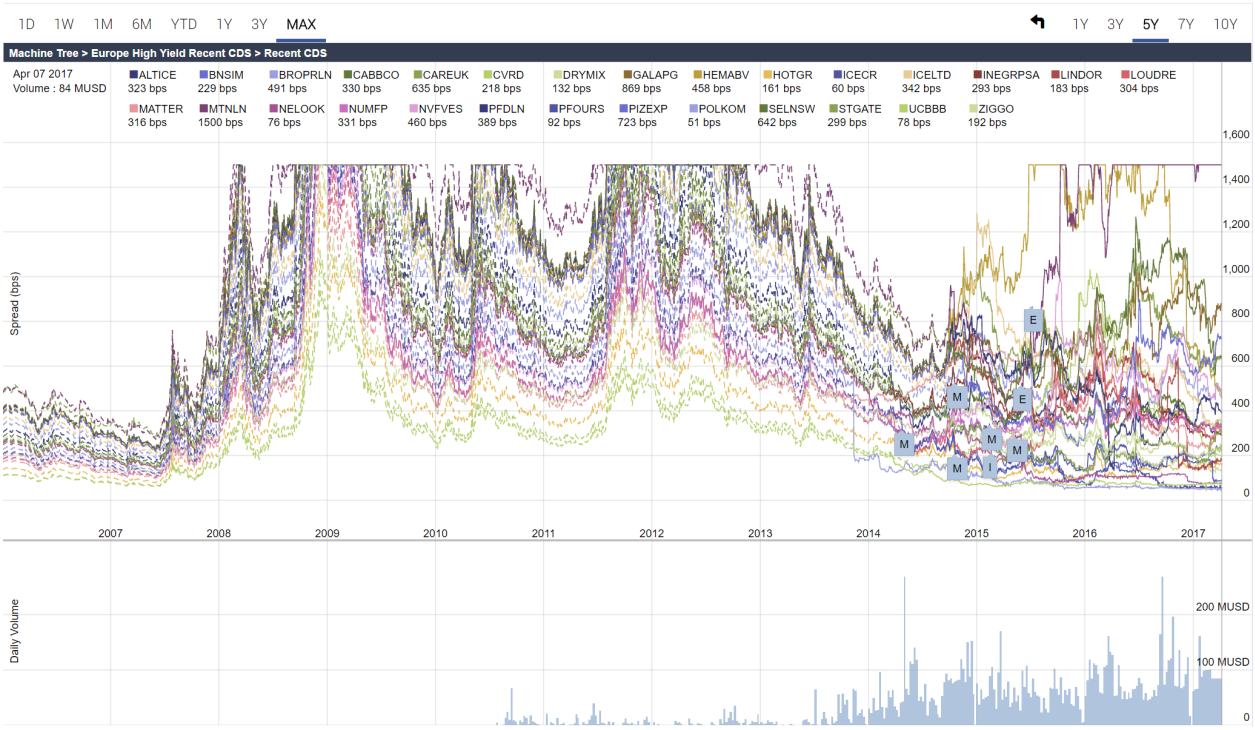


Figure 6.4: Same cluster as in Figure 6.3, i.e. high-yield European entities in the iTraxx Crossover index with partial history. Some of the names have been introduced earlier, for example POLKOM (Eileme 2 AB (Publ)) has been introduced in September 2013, UCBBB (UCB Pharma) has been introduced in March 2014. Their history have been imputed so that the time series start 01/01/2006. We can notice that they share a common history. They are now in the same cluster because their historical data were imputed using the same cluster for the regression. Going forward, with more market data, they may be projected on different clusters (which should be more and more relevant as market prices are obtained) for the regression and thus their imputed past may change, and so their final cluster. Hopefully, it may converge to the 'true' clustering, if any.

6.3 Hierarchical clustering visualization

In this section, we present a visualization that is useful to answer such questions:

- How strong and homogeneous are the clusters?
- How strongly connected / correlated to others clusters are they?
- Is the clustering structure strictly hierarchical?
- Are there overlapping clusters?
- What is the content of a cluster?
- What are the main characteristics of a cluster? of its components?
- Can we name automatically a cluster?

The visualization displayed in Figure 6.5 helps to answer quickly to these questions besides quantitative indicators. It builds upon a technique called seriation (cf. [128] for an historical overview) which aims at reordering the matrix so that its structure appears more blatantly. There are many ways to serialize a similarity (or dissimilarity) matrix and hierarchical clustering is such one. Though hierarchical clustering may not be the 'best' approach possible for the visualization purpose, the same algorithm has to be used to correctly explore the clusters obtained. We already mentioned at the end of Chapter 4 how to leverage the dendrogram to obtain a reordering of the rows (and columns) of the correlation matrix by recursively traversing it from top to bottom using the idea of the quicksort.

In Figure 6.5, we can see the results obtained on historical time series of returns for the STOXX Europe 600. The historical prices are adjusted for all cash and special dividends, splits and all capital changes to produce homogeneous time series (courtesy of Finaltis).

Unlike the hierarchical correlation structure for credit default swaps displayed at the end of Chapter 4 which is quite clear, the one for these stocks is more fuzzy. We can notice that small blocks appear on the diagonal corresponding to subsectors such as, for example, the UK Real Estate Investment Trusts (REITS) depicted in the bottom window having an average correlation of 0.67. The clusters boundaries are overlaid over the correlation matrix heatmap: we can see $K = 26$ clusters which correspond more or less to subsectors which are themselves more or less strongly correlated and more or less interdependent. To see how good a flat partition of K given clusters is, we inspect how their boundaries overlay the coefficients. Do the boundaries cut homogeneous clusters? For example, on the top window, with $K = 2$, we can see that the small cluster does not contain strongly correlated stocks but all these clusters are weakly correlated with those in the big cluster; the big cluster is too coarse, it clearly contains different groups.

Visualization features: When hovering a position (x,y) we obtain the two corresponding stocks and their correlation. If we hover a cluster, we get its components, its mean correlation, and the main characteristics of this cluster based on meta-information for each its stocks (quotation place, industry, sector, subsector, market capitalization, etc.). This information could also be useful for an automatic naming procedure of the clusters (which we used on the HCMapper/Sankeys described in the next section). We note that automatic naming of the clusters is not a well developed research topic and that bespoke recipes prevail though that this issue is a very important practical one.

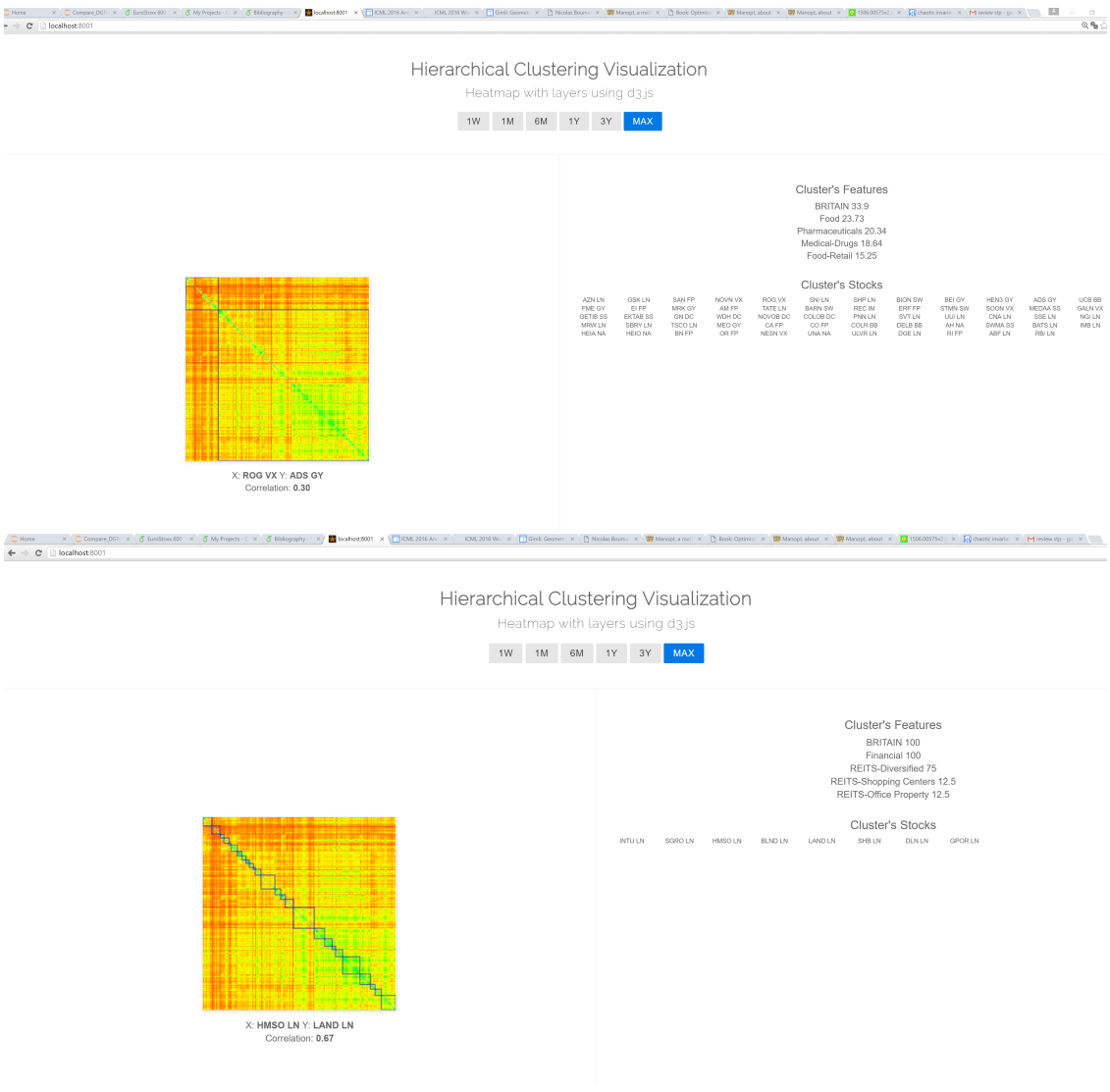


Figure 6.5: Seriation of the STOXX Europe 600 correlation matrix by hierarhical clustering and for hierarchical clustering investigation

6.4 Experimental guidelines to investigate clustering stability

6.4.1 Visualization and comparison of clusters

We describe in this subsection a new visualization tool, dubbed HCMapper, that visually helps to compare a pair of dendrograms computed on the same dataset by displaying multi-scale partition-based layered structures. HCMapper is specifically designed to grasp at first glance both whether the two compared dendrograms broadly agree and the data points on which they do not concur. HCMapper is currently released as a visualization tool on the DataGrapple time series and clustering analysis platform at `www.datagrapple.com`.

Visualization construction

We start from a dataset $\mathcal{X} = \{x_1, \ldots, x_n\}$. We obtain a dendrogram on $\mathcal{X}$ using a hierarchical clustering algorithm. For comparing two dendrograms built over $\mathcal{X}$, we extract from each dendrogram all possible flat partitions over $\mathcal{X}$, thus transforming each one into a tree whose vertices at a given depth define a partition over $\mathcal{X}$, partitions ranging from the coarsest one at the root to the finest one at the leaves as illustrated in Figure 6.6.

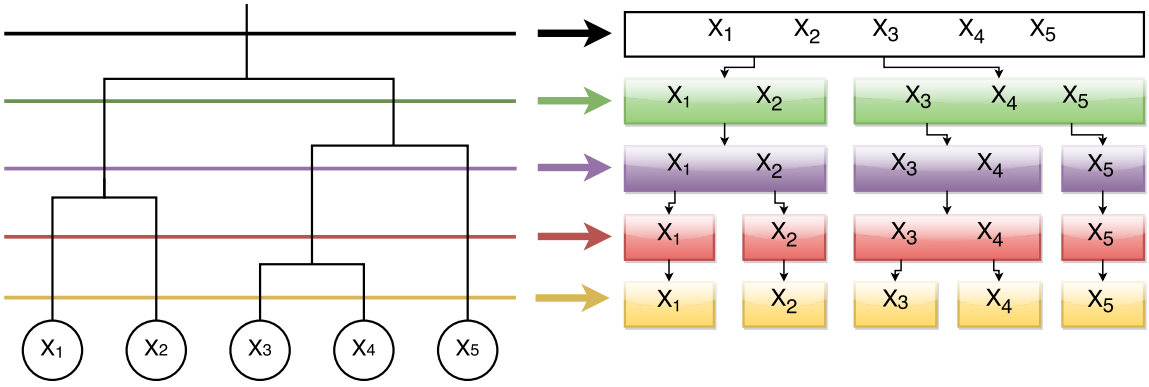


Figure 6.6: Extracting flat partition-based clustering from a dendrogram and transforming it into a tree of clusters; all clusters at a given depth in this tree form a partition over the dataset $\mathcal{X} = \{x_1, x_2, x_3, x_4, x_5\}$.

Bonds between the two tree vertices are explicitly encoded as links representing whether their associated cluster intersects or not. Since cluster size is a significant information of clustering, we display it through its associated vertex whose size is proportional to the ratio of the cluster size over the dataset size. Even more important to our task, understanding clusters intersection: intersection size is encoded through the size of the edge that bonds them, wider is the edge relatively to the cluster size, bigger the intersection; intersection content can be displayed by hovering over the edge. To display this graph, we can leverage the D3.js [24] Sankey, a highly customizable chart, which is amenable for adding application-oriented features.

Concerning time complexity for building the visualization from scratch on $\mathcal{X} = \{x_1, \ldots, x_n\}$, it requires $O(n^2 \log n)$ for applying agglomerative hierarchical clustering algorithm, then for transforming the dendrograms obtained it costs $O(n^2)$, and finally for displaying V vertices $O(V^3)$.

Visualization interpretation

The HCMapper graph is depicted in Figure 6.7. There are three main areas: two contexts, and a focus. The left and right contexts (Tree 1 and Tree 2 in the picture) represent some

of the successive coarser layers of the left and right inner layers (colored layers in the image) to be compared in the focus area. The main goal is to understand how these two partitions are related. In Figure 6.7, we compare a partition at depth n obtained from a dendrogram reflecting hypothesis 1, and a partition at depth m obtained from the alternative dendrogram reflecting hypothesis 2. Looking at the left partition in the focus area, we can see that it is composed of three clusters (green, yellow, orange) of approximately equal size. The right partition consists in four clusters (light green, deep green, light orange and red). We can observe that the left green cluster is mapped onto the light and deep green clusters from the right partition which consist in a refinement. However, right partition is not a strict refinement of the left partition as it can be seen by looking at the right light orange cluster. Indeed, this one is composed of the left yellow cluster and a part of the left orange cluster, the latter evenly splitting into the right light orange cluster and the red one. But, what is the most important to notice here, and which is visually blatant, is the small edge diverging from the bulk of edges linking the left green cluster to the two right green clusters. This singular edge links an element from the left green cluster to the corresponding same labeled element yet belonging to the red cluster in the right partition. This actually can be considered as an outlier or a point of special interest since the two hypotheses broadly agree but on this point. Notice that finding so readily this special point would be much harder by looking through clustering textual results or by inspecting common visual comparison between dendrograms such as the tanglegram displayed in Figure 6.9 which is implemented in the dendextend R package.

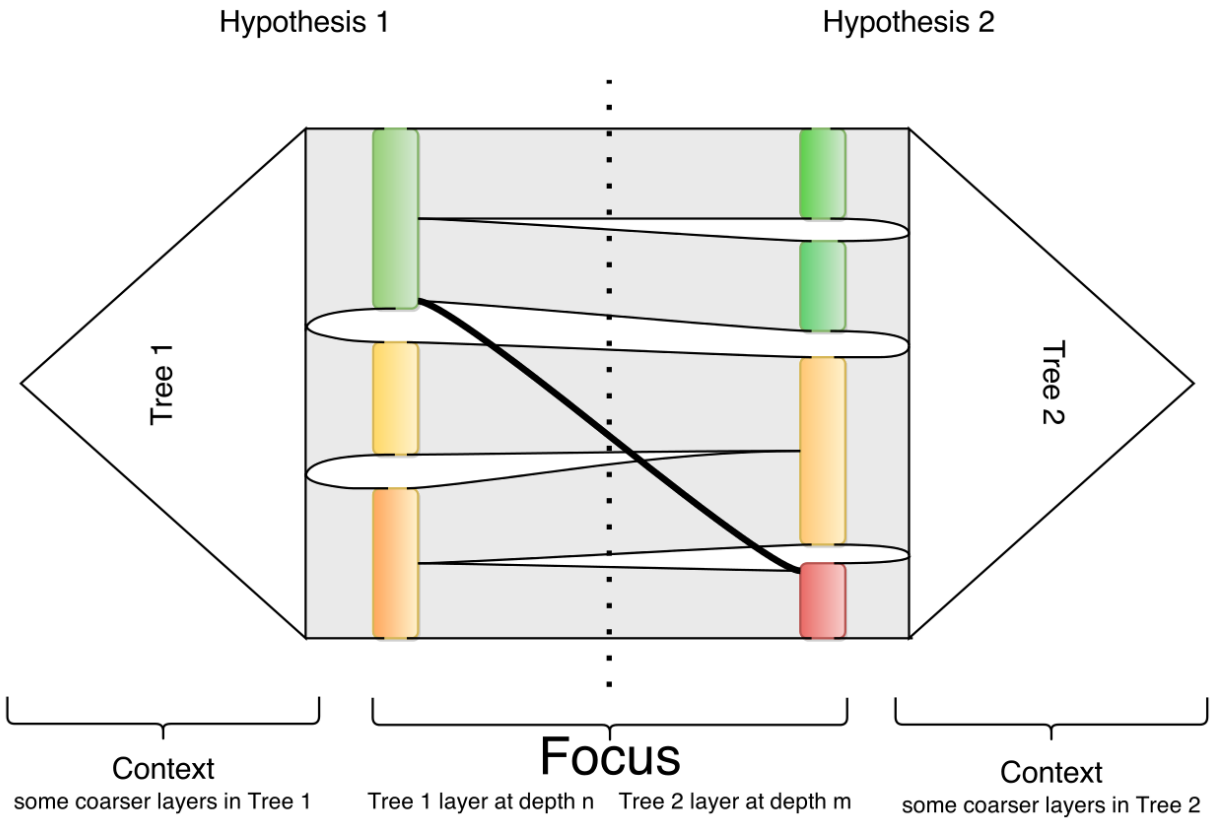


Figure 6.7: Two hypotheses are compared through the dendrograms which were transformed into two trees of partition-layers; note the diverging edge from the green cluster to the red one which highlights a moot point of special interest for experts.

A simple use case: Comparison of the correlation-only vs. the correlation+distribution clustering

We compute two dendrograms: one dendrogram based on correlation only (H1); one dendrogram based on the correlation+distribution distance proposed in Chapter 5 (H2). The resulting HCMapper graph is displayed in Figure 6.8. At once, one can notice that clusters for (H1) are broadly in a one-to-one correspondence with clusters for (H2), but a few outliers highlighted by thin diverging edges. Thus, thanks to the HCMapper visualization, we can conclude that correlation is the main explanatory factor for the clustering of prices time series since only few clusters are modified by adding the distribution information. This can be explained by market microstructure: market makers are specialized and cover specific sectors (as we have seen in Chapter 3), thus adding correlation between prices which are already influenced by common macroeconomic factors. Yet, the few moot points found are of paramount significance for experts as they may correspond to assets whose price variations undergo heavy-tailed distribution or suffer from illiquidity, therefore particular attention should be given to these assets while performing a risk analysis.

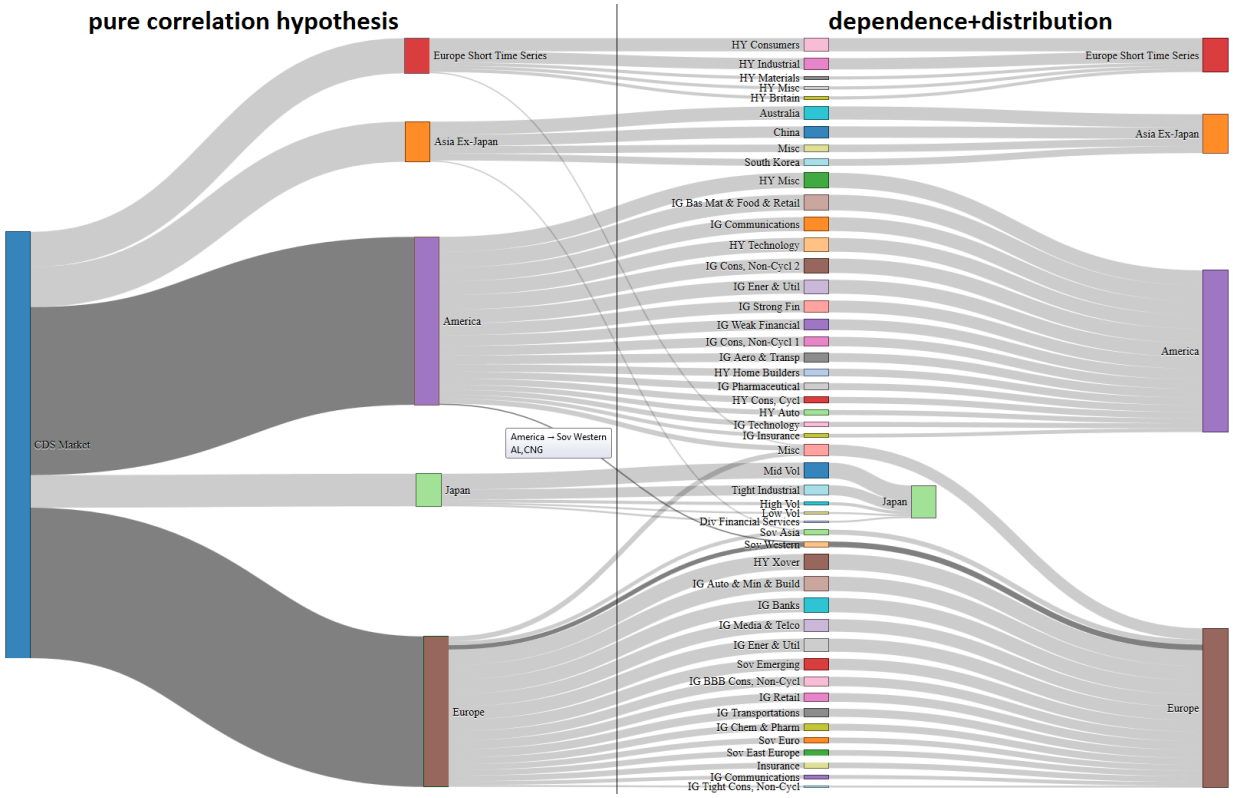


Figure 6.8: Left: two partitions extracted from a dendrogram built from correlation only; Right: two partitions extracted from a dendrogram built using a dependence+distribution distance. By hovering over the thin edge diverging from the “America" cluster in the correlation-only case, we can read AL (Rio Tinto) and CNG (AT&T) on the tooltip. These American companies are clustered with high quality European government debt assets (e.g. Norway, Sweden, Denmark), all of them having a daily variation distribution characterizing illiquid products.

6.4.2 A perturbation framework for testing clusters stability

We present in this subsection an empirical framework motivated by the practitioner point of view on stability. Clustering stability designates the reproducibility of the clustering when data is slightly perturbed [49]. Clustering stability for model selection or validity assessment [124] is indeed a hot topic in the machine learning literature: [14] warn against its irrelevant use as stability only depends on the uniqueness of the clustering objective function minimizer

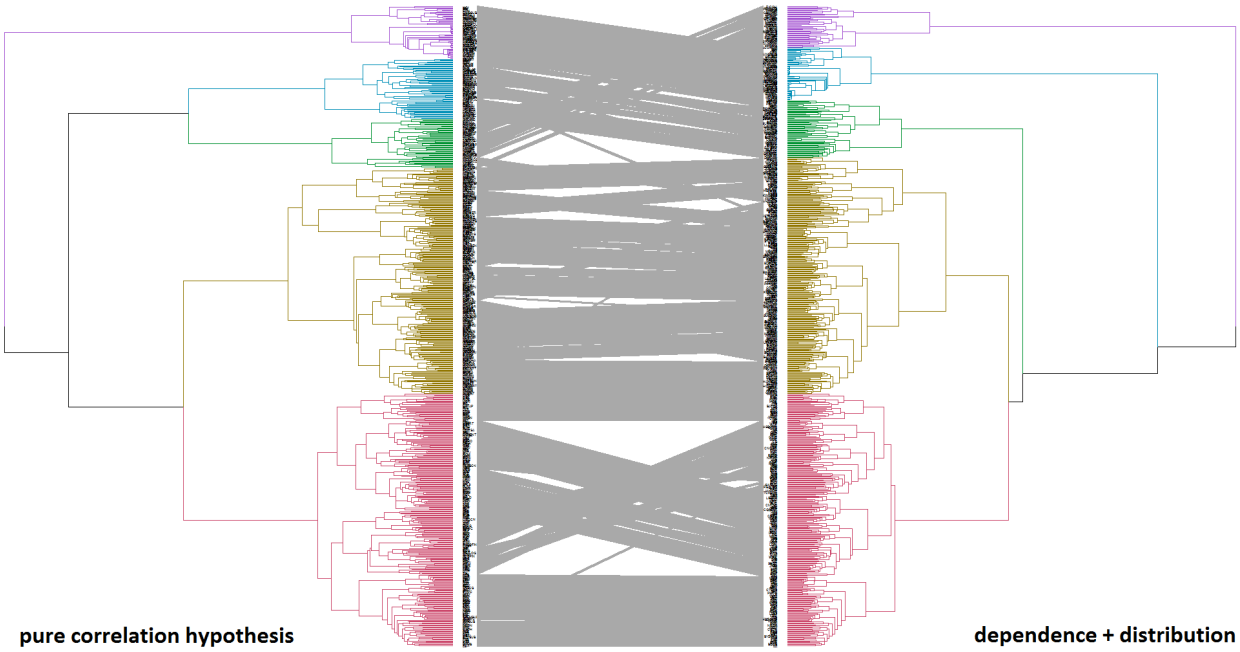


Figure 6.9: Left: two partitions extracted from a dendrogram based on correlation only; Right: two partitions extracted from a dendrogram based on our dependence+distribution distance; Consensus or divergence is much harder to grasp in the dendextend tanglegram [85] than in HCMapper.

for large sample size, yet [189] advocate that this criterion remains useful in the case of finite possibly small samples. From the quantitative analyst or trader perspective, dynamical stability of models, i.e. stability to online perturbations arising from streaming data, is required for being confident using them. A clustering should only change when a meaningful event happens in the market. Despite being an important notion for practitioners, few works have dealt with the dynamical stability properties of clusters computed on financial time series [64].

The goal of our perturbation framework is to both assess clustering validity and yield market insights by providing through the data perturbations we propose a multi-view of the assets' clustering behaviour.

Perturbations can be performed both on returns (some of the T values of each time series) or on assets (some of the N time series themselves). Concretely, these perturbations consist in modifying row-wise or column-wise the $N \times T$ data matrix X.

We provide below a list of perturbations concerning some of the T time series values. We explain their motivations arising from financial concerns and what we can learn from them by analyzing the clustering stability.

Sliding Window

Motivation: Dynamical stability of models is a requirement for trading and risk information systems. For example, value at risk (VaR), an estimated amount of money so that the potential loss of a portfolio over a given timespan should not exceed, is computed on a moving window and updated with respect to the asset prices stream. With no trading in the portfolio, and in a stationary regime, VaR should not vary too much.

Definition: Given a window width W and a step size S, clustering is performed on $X_{:,[t_{\text{cur}},t_{\text{cur}}+W[}$ and $X_{:,[t_{\text{cur}}+S,t_{\text{cur}}+S+W[}$, then current time t_{cur} is updated $t_{\text{cur}} := t_{\text{cur}} + S$, and so on.

Insight: A clustering that strongly differs from one time to another when the market seems in a steady regime should be rejected since very sensitive to noise, i.e. small unsignificant market variations. If confidence is high in the methodology, modification of clusters may be

a signal that market structure is changing (for instance, end of a crisis and decrease in global correlation).

Odd vs. Even

Motivation: A clustering algorithm applied on two samples describing the same phenomenon should yield the same results. How to obtain two of these samples? The goal is to split the sample in two while mitigating the effect of non-stationarity, seasonality, end-of-the-week trading activity, meetings and announcements from the ECB or the FED generally happening on Friday.

Definition: We define $X^{(1)} = \{X_{:,t} \mid t \text{ is odd}\}$ and $X^{(2)} = \{X_{:,t} \mid t \text{ is even}\}$, i.e. we build the sample of the odd trading days and the sample of the even trading days. Since the trading week lasts 5 days, we alleviate the aforementioned statistical biases. Clustering is performed independently on $X^{(1)}$ and on $X^{(2)}$.

Insight: If not stable, the clustering method should be rejected.

Economic Regimes

Motivation: Since the economic context can change dramatically, financial time series do not evolve in a steady regime in the long run. It makes sense to split the timespan into different periods where the statistical regime can be considered stationary.

Definition: We partition the sample into M subsamples $X = \sqcup_{i=1}^{M} X^{(i)}$, where $X^{(i)} = X_{:,[t_i,t_{i+1}[}$ and the $(t_i)_{i=1}^{M+1}$ delimit time intervals. The breakpoints $(t_i)_{i=1}^{M+1}$ can be chosen guided by market understanding or computed using a dynamical changes and regime detection algorithm.

Insight: We can study whether the clustering structure is persistent throughout different economic regimes, and how strongly it is. If not, which are the periods concerned and how steep is the change? Which assets are involved? Why do they switch from clusters? What happened to them and their clusters? However, we must keep in mind that it is difficult to separate the signal from the noise of the clustering methodology.

Heart vs. Tails

Motivation: Does the market under stress share a common clustering structure with the market during uneventful periods?

Definition: Let $\overline{X}_t = \frac{1}{N}\sum_{i=1}^{N} X_{i,t}$ be the mean time series of the market. let Q_1 be the lower quartile and Q_3 be the upper quartile. We define $\mathcal{T} = \{t \mid \overline{X}_t \leq Q_1 \vee \overline{X}_t \geq Q_3\}$ and $\mathcal{H} = \{t \mid \overline{X}_t \notin \mathcal{T}\}$ corresponding to times having market values in the tails and in the heart respectively. Then we split the sample X in the two following subsamples $X^{(1)} = \{X_{:,t} \mid t \in \mathcal{T}\}$ and $X^{(2)} = \{X_{:,t} \mid t \in \mathcal{H}\}$ on which we apply the same clustering algorithm.

Insight: Although it is difficult to anticipate changes of the market behaviour, in period of stress all assets tend to be simultaneously affected by macroeconomic tensions which usually induce a significant increase in correlation betweem them (cf. Fig. 6.10). Thus, correlation should be less discriminating and a correlation-based clustering might be unstable with respect to this perturbation.

Multiscale

Motivation: Markets prices can be monitored from a high frequency sampling (tick by tick or minute by minute) to much lower frequency (from hours to hours, days by days or on a weekly basis). The sampling frequency used is linked to the type of trading, from high-frequency trading (HFT) and algorithmic trading to long-term investments. Is the clustering structure persistent throughout a wide range of time scales or does it strongly depend on the sampling?

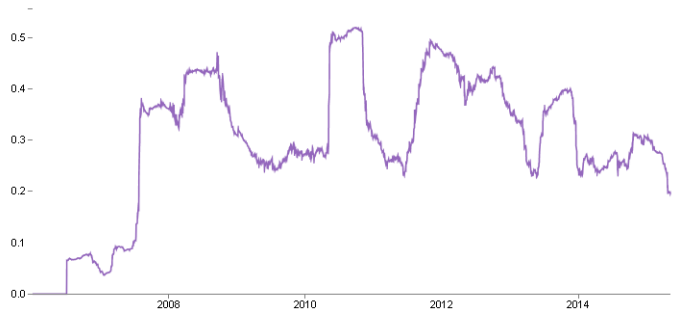


Figure 6.10: Mean Correlation Dynamics computed on the whole CDS dataset from 2006 to 2015 using a 6-month sliding window

Definition: From X we can build M datasets $X^{(i)} = X_{:,s_i}$, $i = 1, \ldots, M$ where s_i are regularly-spaced multiscale subsample of $\{1, \ldots, T\}$.

Insight: Ideally, in the perspective of building a risk system, we would like that the choice of risk factors is independent of the time scale used for the analysis. This perturbation allows us to verify to which extent clustering is multiscale persistent.

Maturities

Motivation: Fixed-income assets such as bonds, swaps and CDS for instance, have a lifespan called their maturity. Several products with different maturities (say an insurance against the default of a corporate for 1, 2, 3, 5, 7 or 10 years) can concern the same entity. Since the underlying risk is the same, we would like similar clusterings.

Definition: We get from the market several time series dataset $X^{(i)}$, one $N \times T$ data matrix for each quoted maturity.

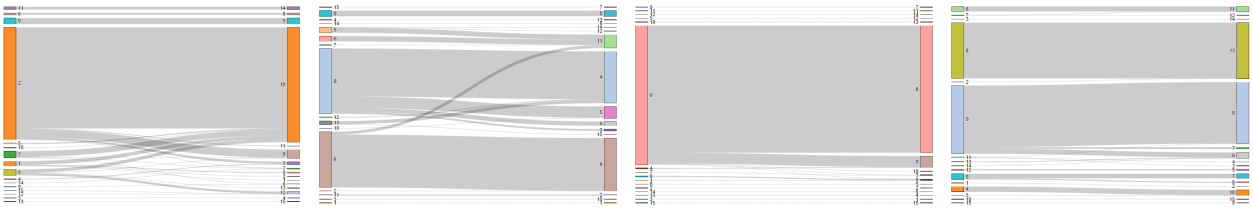
Insight: We can either reject a clustering method which yields to unstable clusters or investigate why a particular maturity has a different clustering structure compared to the others.

Term Structure

Motivation: The term structure is the set of all quoted maturities. Clustering term structure could lead to a more meaningful result than clustering separately each maturity.

Definition: For clustering term structures, one need a specific distance. To our knowledge, the problem of obtaining a proper one which captures the whole information (e.g. dynamics, distribution and correlation of its distortions, intra- and inter-correlation) has not been addressed. Here, we give a simple one for clustering CDS term structures at a given date t. A CDS probability of default $P(t)$ can be viewed as a cumulative distribution function on $\mathbf{R}^+$. Indeed, the probability of default is increasing, the probability of instantaneous default is 0, and at infinity all entities will eventually default. Thus, $f(t) = \partial P(t)/\partial t$ defines a probability density function on $\mathbf{R}^+$, and since $\int_{\mathbf{R}^+} f(t)\ dt = 1$, $\sqrt{f(t)}$ is a unit vector in $L^2(\mathbf{R})$. The inner product between two unit vectors defines an angle ϕ which is the distance between two term structures. Given two term structures P_1, P_2 and f_1, f_2 such that $f_i(t) = \partial P_i(t)/\partial t$, their distance ϕ can be written $\cos\phi = \int_{\mathbf{R}^+} \sqrt{f_1(t)}\sqrt{f_2(t)}\ dt = 1 - H^2(\sqrt{f_1(t)}, \sqrt{f_2(t)})$, where H is the Hellinger distance.

Insight: For entities near default, the term structure should be inverted, i.e. the market anticipates a renormalisation if these entities survive. For entities having seemingly no troubles, the quoted term structure should mirror the debt term structure of these entities. Some industries have a particular debt structure (short term debt for financials, long term debt for basic materials and industrials). Part of this information should also be captured by correlation between assets on a given maturity.



D^P ARI 0.46 D^S ARI 0.71 D^E ARI 0.47 D^G ARI 0.90

Figure 6.11: Stability to Odd vs. Even perturbation and the associated ARI showing a better stability of D^G-based clustering; partitions obtained from D^S-based clustering are rather similar but less stable

To the presented time-based perturbations, we add the following two population-based perturbations on the set of assets: **increasing/decreasing the number of entities** and **adding entities with imputed historical prices**. These perturbations can be easily motivated: new companies emerge regularly and some others disappear from the market. The clustering structure should not radically change when adding or removing entities from the clustering perimeter. When new companies are created and introduced in the market, they have not much history. It may be necessary to impute missing data based, for instance, on a clustering methodology. We would like to verify that adding synthetic time series built from existing ones to the clustering perimeters does not change the original clustering structure. The clustering structure should be robust to the statistical engineering performed to impute missing data or clean their poor quality.

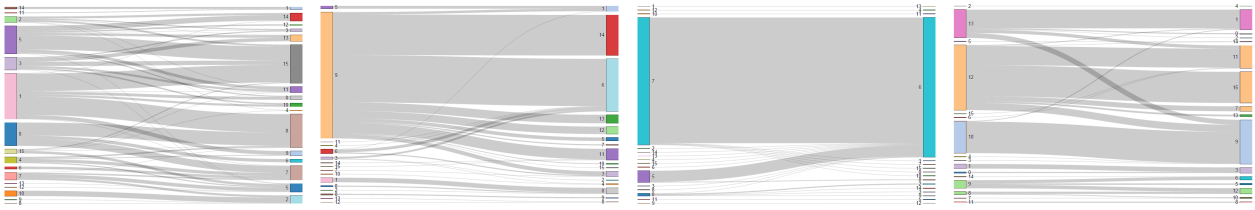
This list of perturbations is obviously not exhaustive but gives an idea on how to empirically investigate the clusters. For example, one could think of using different CDS datasets of historical prices and compare the results since these datasets differ in quality as pointed out in [146]. It would be embarrassing for policymakers or traders that their decisions are 'overfitted' to some measurement noise and carefree recording of data.

Comparison of distances using the perturbation framework

We leverage the proposed perturbation framework to test four distances used for clustering financial time series. We also observe some stylized facts about the CDS market. The four distances D^P, D^S, D^E and D^G are essentially distances based on Pearson correlation, Spearman correlation, Euclidean distance and the distance introduced in Chapter 5 based on correlation and distribution respectively. We illustrate the clustering stability with the HCMapper visualization, essentially a Sankey diagram, which highlights the dissimilarties between partitions as explained in the previous subsection.

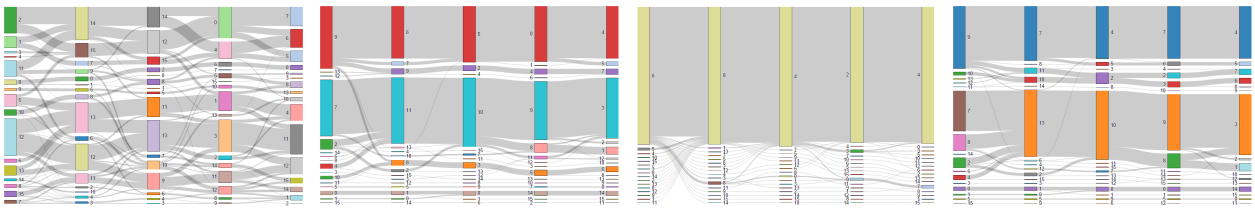
In Fig. 6.11, we display the stability results on the Odd vs. Even experiments. For each distance, we have displayed the partitions obtained on the odd trading days sample (left) and on the even trading days sample (right). A grey link binds a given asset in the left partition to the same asset in the right partition. Thus, a perfectly stable clustering is displayed by a one-to-one correspondence between left and right clusters. Diverging edges highlight mismatches between partitions, hovering on the edges shows the assets switching from clusters. In this experiment, these can be assets with an unusual history, for instance they may have encountered a strong variation on a particular day due to a merger (M&A), a catastrophe or a fraud. But, of course, a cluster switch can happen due the clustering method shortcomings.

The Heart vs. Tails experiment displayed in Fig. 6.12 shows an interesting stylized fact about the CDS market. Clustering on correlation (D^S and D^P) is not stable at all with respect to this perturbation. This means that the sample of the strongest moves in the market has a totally different clustering structure than the sample of the mildest moves when considering only correlation. This can be explained since when the market is stressed, macroeconomic tensions tend to affect all the participants and correlation between assets be-



D^P ARI 0.24 D^S ARI 0.09 D^E ARI 0.31 D^G ARI 0.43

Figure 6.12: Stability to Heart vs. Tails perturbations for different distances and the associated ARI



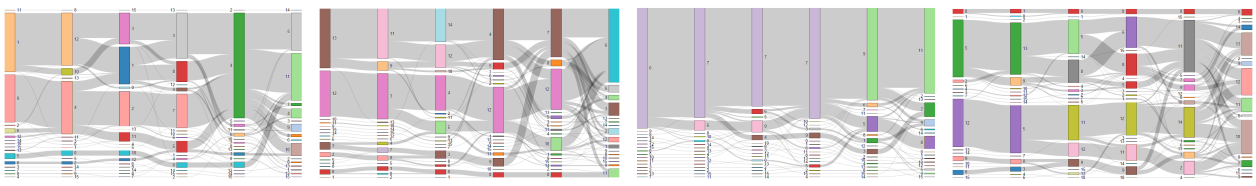
D^P		D^S		D^E		D^G	
$1 \to 3$	0.34	$1 \to 3$	0.71	$1 \to 3$	0.05	$1 \to 3$	0.61
$3 \to 5$	0.43	$3 \to 5$	0.92	$3 \to 5$	0.62	$3 \to 5$	0.90
$5 \to 7$	0.50	$5 \to 7$	0.86	$5 \to 7$	0.61	$5 \to 7$	0.83
$7 \to 10$	0.46	$7 \to 10$	0.88	$7 \to 10$	0.87	$7 \to 10$	0.87

Figure 6.13: Stability to Maturity perturbations

comes significatively higher and similar for all assets, thus becomes uninformative. This claim is supported by the fact that D^S based on the Spearman correlation (correlation between ranks) performs the worst, whereas D^P, based on the Pearson correlation measure known to be decreased by fat-tailed variations, achieves a better stability since this correlation-based distance discriminates unintentionally on distributions. For high values of ρ, which is the case in stressed period, D^E discriminates on the mean and variance of the variations, so performs better than the correlation-based distances. Finally, D^G which intentionally works on both information can leverage the distribution information and obtain a rather stable clustering between the stress periods and the more quiet ones.

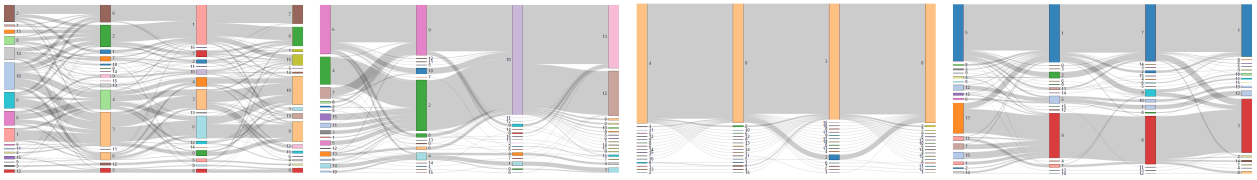
In Fig. 6.13, we display results of the Maturity experiment. For each clustering, we show 5 partitions corresponding to clustering the 1,3,5,7,10-year CDS. We can notice that the partition corresponding to the 1-year CDS is the less stable whatever the distance used. This can be explained by the relative illiquidity of the 1-year maturity compared to the others yielding to scarce and noisy quotes from the market makers. Stability is high for D^S and D^G and abnormally low for D^P and D^E while information is essentially the same.

The Fig. 6.14 depicts results of the Multiscale experiments. 6 partitions are displayed for each distance corresponding to the clusterings obtained by considering respectively



D^P		D^S		D^E		D^G	
$1 \to 2$	0.81	$1 \to 2$	0.86	$1 \to 2$	0.42	$1 \to 2$	0.81
$2 \to 4$	0.41	$2 \to 4$	0.62	$2 \to 4$	0.63	$2 \to 4$	0.58
$4 \to 8$	0.58	$4 \to 8$	0.55	$4 \to 8$	0.59	$4 \to 8$	0.61
$8 \to 16$	0.12	$8 \to 16$	0.30	$8 \to 16$	0.43	$8 \to 16$	0.48
$16 \to 32$	0.07	$16 \to 32$	0.15	$16 \to 32$	0.60	$16 \to 32$	0.11

Figure 6.14: Stability to Multiscale perturbations



D^P		D^S		D^E		D^G	
ARI	0.13/0.30/0.30	ARI	0.36/0.14/0.16	ARI	0.26/0.02/0.10	ARI	0.40/0.69/0.60

Figure 6.15: Stability to Economic Regimes perturbations

1,2,4,8,16,32 trading days variations. We can observe as a stylized fact that the clustering structure is persistent up to a weekly sampling, and that the clustering structure is essentially determined by correlations as advocated by the high stability achieved by D^S and D^P. D^G, once again, is relatively stable leveraging its correlation part which is similar to D^S.

We finally conclude this empirical study with the Economic Regimes perturbations. In Fig. 6.15, we display 4 partitions corresponding to the clusterings obtained on different economic periods. From left to right, the pre subprime crisis period 2006-2007, the subprime crisis period 2008-2009, the European debt crisis 2011-2012 and the quantitative easing 2013-2014. We can notice in Fig. 6.15 that the period 2006-2007 yields very different clusters compared to what follows. Indeed, looking at Fig. 6.10, we observe that correlation in the market was very low. Except clustering with D^G, clusters obtained with the other methods are not stable. The partitions and their stability scores obtained from the D^G-based clustering agree with previous remarks: pre-crisis period was much different, the clustering structure is the same during both crises, and now that correlation is decreasing and that quantitative easing is at work the clustering structure of the market is changing.

6.5 Monitoring clusters

From day to day, the clustering may change. We already explained that some of these changes may be artifacts of the methodologies. Some changes may be of interest. We use the HCMapper to spot these changes (cf. Figure 6.16) and matrix serialization (re-ordering of the rows/columns to make the hierarchical sub-structure appear more blatantly) to inspect the similarity/correlation matrices of the clusters involved in the moves (cf. Figures 6.17, 6.18, 6.19, 6.20). For example, in the case of the 'Korean' cluster displayed in Figure 6.17, HANABK (Korea Exchange Bank (KEB) Hana Bank) joins an already existing cluster of Korean entities (but for the less correlated one, ADGB (Abu Dhabi)). This move may be only due to an oversegmentation of the time series over a short period (1 year) as in the long run, the 'Asia Ex-Japan'/'Korean' cluster is usually and neatly divided into two subclusters: 'Corporate' and 'Financial' (cf. the Machine Tree on DataGrapple `www.datagrapple.com`). Over a short period, they can mix as seen in Figure 6.17: Corporate entities (POHANG (Posco), SAMSNG (Samsung Electronics), KORELE (Korea Electric Power Corporation)), and Financial ones (CITNAT (Kookmin Bank), EIBKOR (The Export-Import Bank of Korea), KDB (The Korea Development Bank), HANABK (KEB Hana Bank), INDKOR (Industrial Bank of Korea)). On the contrary, Figures 6.18, 6.19, 6.20) may be cases of undersegmentation: One can notice several subclusters which are only mildly correlated together. It is not surprising then that these subclusters can split.

We can also monitor the clusters using lots of relevant indicators, but this is much more of an art reserved to practitioners who have particular ideas in mind than science.

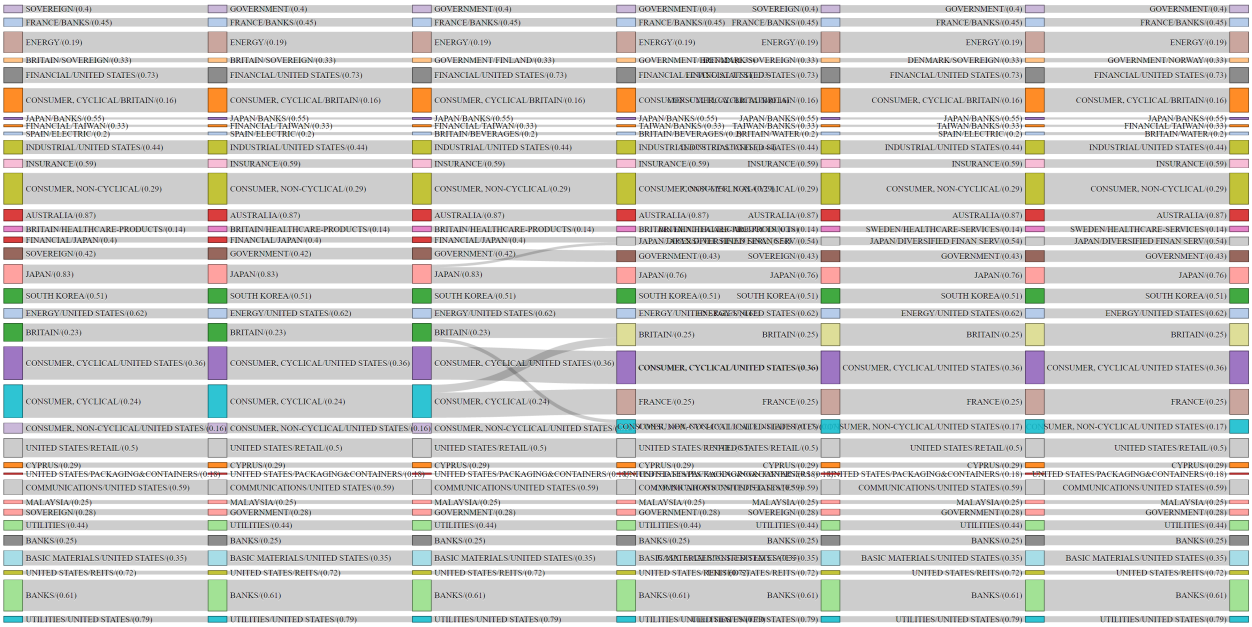


Figure 6.16: The rightmost partition corresponds to the partition as of 'today' (31-03-2017). From right to left, they correspond to the 'today'$-i$, $0 \leq i \leq 6$, trading days. We can notice that 3 trading days ago, there have been some changes in the clusters (e.g., some entities left the 'JAPAN' cluster to move into the 'FINANCIAL/JAPAN' cluster).

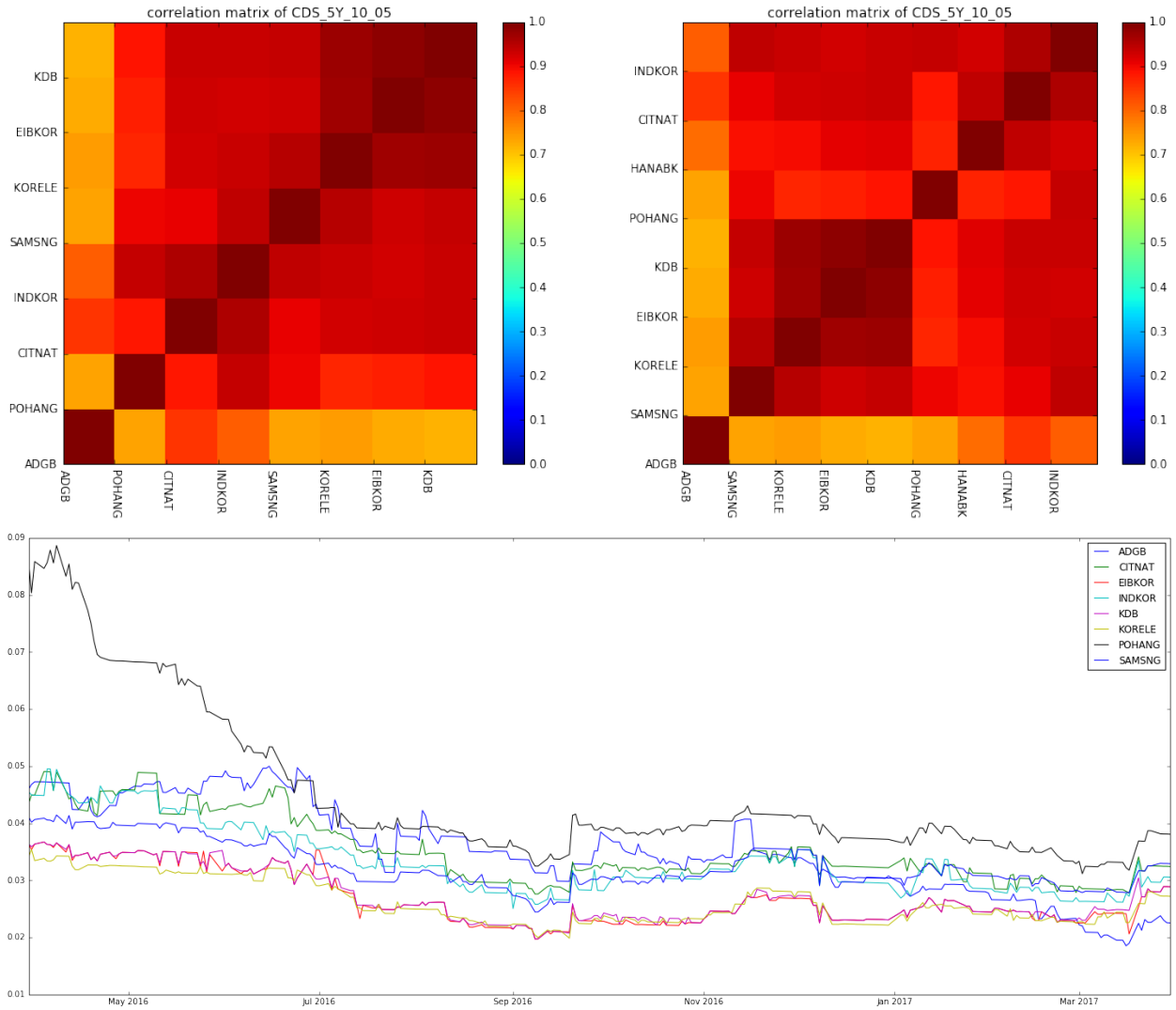


Figure 6.17: Top: Korean cluster at time $t-1$ (left) and t (right): HANABK (KEB Hana Bank, a Korean bank holding) have joined a cluster of Korean companies; Bottom: NPVs

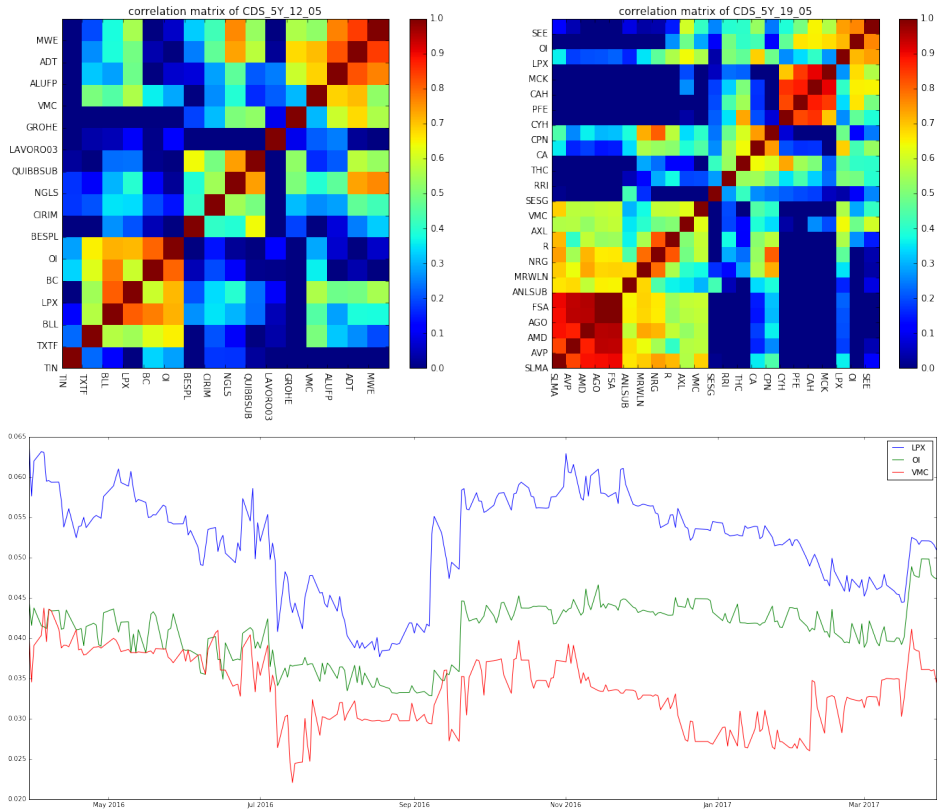


Figure 6.18: LPX (Louisiana-Pacific Corporation), OI (Owens-Illinois), VMC (Vulcan Materials Company) provide building materials. They leave a cluster of industrials which explodes (cf. also next figure) for another one where they seem better integrated.

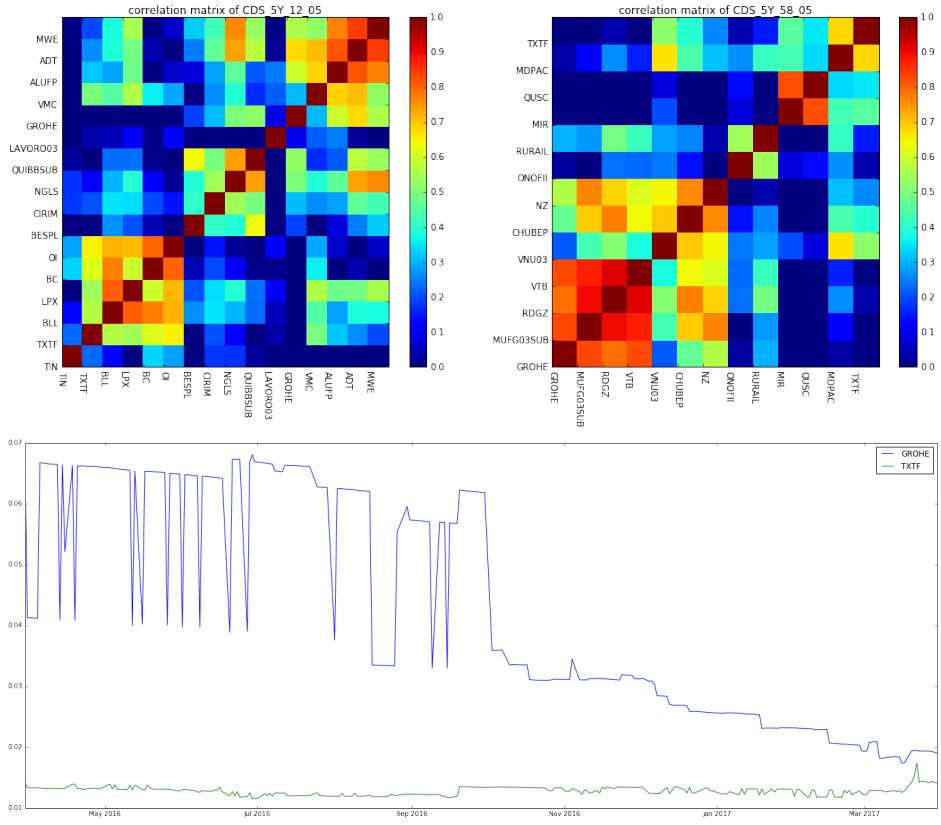


Figure 6.19: TXTF (Textron) is an aerospace, defense, security and advanced technologies conglomerate. GROHE (Grohe) is a sanitary fittings manufacturer. They also move to another cluster which contain mostly illiquid names.

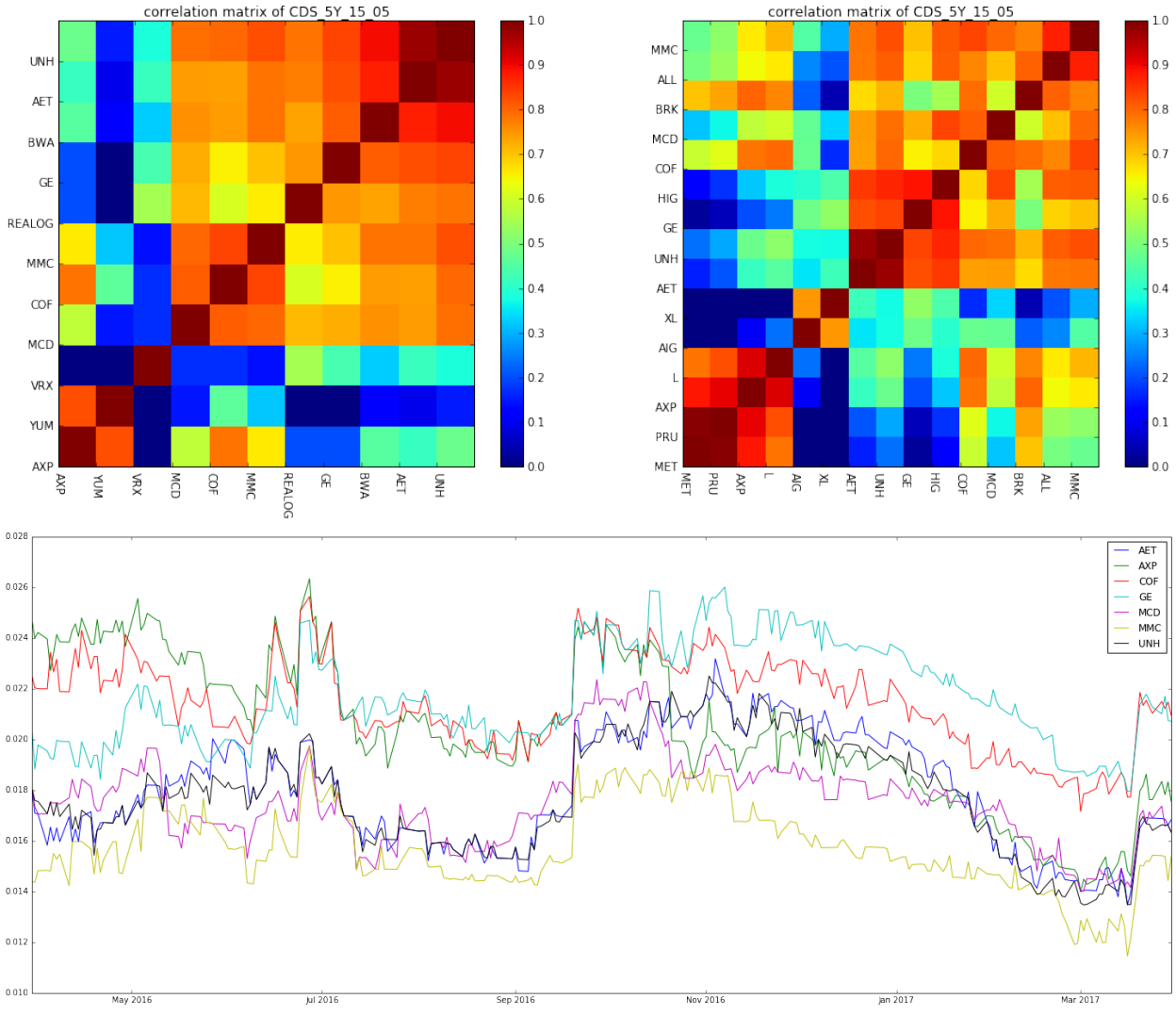


Figure 6.20: AET, AXP, COF, GE, MCD, MMC, UNH (blue chip companies) are joined by ALL, BRK, HIG, XL, AIG, L, PRU, MET (financial services and insurance) and are left by BWA, REALOG, VRX, YUM (misc. industries)

Chapter 7

Conclusion and perspectives

7.1 Summary of contributions and main ideas

In this thesis we have first gathered and reviewed the scattered literature about clustering financial time series (see also [138]). This research topic really emerged two decades ago (in 1999) with Mantegna's paper [133] though some early attempts were already published in the 1960s [112]. This approach, which is essentially computational, benefited from the increase and widespread of computing power in the late 1990s to find a broader public. We have highlighted in this review that most of the papers were focusing on exploration and explanation of the minimum spanning tree and the dynamics of the hierarchical structure, but not on the robustness of these techniques to build information systems on top of them. At Hellebore Capital, clusters are used as building blocks to build risk measures (improved Value-at-Risks and SPAN-like methods) and trading systems. They are also used for data quality: cleaning historical databases, monitoring and finding outliers in the livefeed. We provided in Chapter 6 some practical methodologies to apply clustering analysis on financial time series. Since their applications can be critical, one may want to verify that the methodology is sound hence the consistency part of the thesis (see also [139]). The second and core part is dedicated to distances between financial time series which are designed to increase relevance and robustness of the existing clustering methods with the idea that *a proper distance working on a proper representation of the data should help stabilize the clustering results* (see also [68, 140, 136, 144, 141, 143]).

In Chapter 3, we spent several pages explaining the CDS datasets to highlight the many possible sources of noise: prices are sent by different sources which are more or less reliable; the messages can contain typos; their information is not always clearly structured; the messages can contain ambiguities and implicit information; the 'smart parsing' technology [200] is robust but does not yet capture 100% of the information; we do not necessarily get information from all the current CDS traders; several arbitrary choices have been made such as the rules for a synthetic order book, i.e. the time persistence for a quote or its priority; the snapshot time for building the daily time series, etc. It all results in a very noisy process having possibly many biases. In these conditions, modeling sophistication (e.g. learning models with millions of parameters to fit on small noisy datasets) may have pitfalls and can 'overfit' the noise by recovering statistical patterns which are only artifacts of the data sources or processing pipeline. For example, even with simple clustering algorithms one can discover plainly wrong patterns such as the specific curves of the Japanese CDS term structures which are essentially due to the relatively high frequency of 5-year quotes vs. the lack of frequent information on other maturities (as we have seen in the Chapter 3 volumetry statistics). The number of public holidays (higher in Japan) can also bias the distribution of returns and correlations. We have striven to achieve an appropriate balance in our modeling.

7.2 A research program

Many questions have been raised during our investigation. Not all were answered, and not all that have been answered have been included in the manuscript. We list below issues that we think are important to solve for advancing the research in financial time series clustering:

- **Riemannian geometry of correlation matrices** (Fisher-Rao distance and mean)
  potential use case: moving average of correlation matrices; computing information ball [158] of correlation matrices; study the evolution of the correlation structure inside a given cluster / between the cluster representatives;

- **multivariate correlation** (dependence measure for several random variables/vectors and dependence measure between two random vectors)
  potential use case: portfolio diversification; alternative hierarchical agglomerative clustering algorithm based on the principle that merging groups of variables/vectors should yield a minimum decrease in the resulting dependence of the groups; dependence between random vectors is not well understood (copulas are not suitable anymore, cf. [87], some tools have been proposed [126, 125, 25] but they have not received much follow-up until now); These dependence measures between random vectors would help to leverage richer time series representations of assets;

- **augmented representation** (several time series are used to describe an asset)
  potential use case: using a (credit default swap spread, stock price) time series representation for entities, one can take into account both the credit and equity of the firms into the analysis; using a term structure representation, one can discover entities with a particular debt profile; one can investigate the added value of a bid/ask representation, an open-high-low-close representation, etc.; Which time series should one include in the representation? Having a richer representation, i.e. several time series describing an entity, raises problems: It is harder to understand and measure dependence between random vectors. Developing such tools could:
  - improve stability of the clustering methodology;
  - improve convergence rates of the clustering algorithms (empirical ergodicity);
  - provide an elaborate analysis.

- **number of clusters** (most probably there isn't a 'correct' number of clusters since the structure is highly hierarchical with many stable layers, so an appropriate number has to be task and goal specific)
  potential use case: filtering of covariances; portfolio diversification; cluster-based trading strategies

- better understanding of **entities switching clusters**: noise or signal?
  potential use case: trading signal for trend following or mean reverting strategies.

- obtain more precise results for (empirical) **convergence rates**, and propose a robust tool for checking whether the statistical conditions are good enough for a clustering algorithm to be applied on the financial time series; if not, suggest a minimum sample size.

- build an **open source library of the most common techniques** described in the survey. It can allow a more effective technology transfer from academics to practitioners and it can also help researchers to compare more efficiently the methodologies they suggest without implementation biases.

- create and provide **synthetic datasets** (essentially paths of random processes) **containing stylized facts of financial time series** and acting as gold standard benchmarks to understand and compare the behaviour of the different methods.

Résumé de la thèse

Introduction

La modélisation statistique des séries temporelles financières est une tâche difficile, mais peut permettre une meilleure gestion du risque et des investissements. Cette tâche est difficile car les séries temporelles financières exhibent des comportements particuliers : non-stationarité, rapport 'signal / bruit' faible, peu d'observations par rapport au nombre de variables à étudier. Pour ces raisons, calibrer des modèles robustes qui sont également valables sur des données futures est compliqué. La plupart des modèles ne fonctionnent que pour une certaine période et fournissent par la suite des résultats trompeurs. Le clustering, qui peut permettre de regrouper des actifs ayant un comportement similaire, aide à réduire la dimensionnalité des données et peut constituer une brique de base importante dans l'élaboration de modèles robustes.

Dans cette thèse, nous nous efforcerons de justifier la construction et l'utilisation des clusters comme pré-traitement utile à la modélisation des séries temporelles financières.

Des données et un état de l'art épars

La littérature sur le clustering de séries temporelles financières est répartie dans différents domaines : la physique statistique et l'économétrie, le data mining et le machine learning, la comptabilité et la finance. Ces domaines communiquent relativement peu, et les conclusions respectives ne sont pas encore organisées clairement pour pouvoir constituer un socle solide de connaissances.

Le marché de gré à gré des couvertures de défaillance (en anglais, credit default swaps) a cela en commun. La donnée (prix, volumes, transactions effectuées) est répartie dans de nombreuses bases qui ne se concilient pas facilement, elle est coûteuse à acquérir et à maintenir.

État de l'art

La majorité des études se concentre sur l'utilisation de l'arbre couvrant de poids minimal construit à partir d'une estimation des corrélations linéaires (estimateur de Pearson) pour étudier la structure hiérarchique des corrélations pour tel ou tel marché. Il est maintenant connu que la plupart des marchés financiers vérifient cette propriéte. Ces études conduisent ensuite à diverses interprétations économiques des résultats (clusters, statistiques descriptives du graphe) et de leurs évolutions temporelles. Relativement peu d'entre elles visent à améliorer la démarche statistique proposée par Mantegna en 1999, et encore moins à proposer les clusters comme briques de base à des systèmes de risque ou de gestion.

Dans cette thèse, nous proposons une revue de la littérature [138] qui essaye de couvrir le plus large spectre possible sur le clustering de séries temporelles financières, allant de leurs conceptions à leurs utilisations.

Les données sur les couvertures de défaillance

Les données utilisées dans cette thèse sur les couvertures de défaillance viennent essentiellement de deux sources : des messages envoyés par les principaux teneurs de marché, et les transactions reportées dans des registres spéciaux dont la création a été imposée par le "Dodd-Frank Act" (21 juillet 2010) suite à la crise financière mondiale de 2007-2008 où l'opacité de ces marchés a joué un rôle.

Dans cette thèse, nous décrivons les traitements de ces données qui ont permis de créer les séries temporelles de spreads des credit default swaps sur lesquelles nous avons ensuite travaillé. Le fonctionnement de ce marché, et les procédures de traitement des données [200], ont très certainement ajouté du bruit supplémentaire au bruit structurel de marché.

De la consistance du clustering

Pour que les praticiens aient davantage de confiance dans ces méthodes de partitionnement automatique des données (clustering), il est nécessaire de leur fournir certaines garanties. La consistance en est l'une d'entre elle : pour peu qu'il y ait suffisamment de données, est-ce que le modèle trouve toujours les bons résultats ? Dans le chapitre de cette these dédié à cette question, nous montrons de tels résultats, sous certaines hypothèses [139]. Nous nous intéressons ensuite à étudier empiriquement les convergences empiriques des méthodes. Pour le praticien, en effet, il est non seulement important d'utiliser des méthodes fondées, mais encore faut-il qu'elles puissent fonctionner dans les conditions auxquelles il est confronté, c'est-à-dire relativement peu d'historique utilisable de manière pertinente, et beaucoup de bruit. Nous cherchons donc des méthodes robustes et qui convergent vite vers les résultats attendus. L'étude empirique permet d'en exhiber quelques-unes.

De nouvelles distances entre séries temporelles corrélées

La seconde partie de cette thèse part du constat que les distances utilisées dans les méthodes de clustering décrites dans la littérature sont en général simplistes et fondées sur l'utilisation de la corrélation linéaire qui ne capture qu'une très faible partie de l'information disponible. Dans ce chapitre, nous proposons deux nouvelles distances qui permettent de pallier certains de ces problèmes.

Une distance en corrélation et distribution

La littérature se concentre sur la corrélation, les comovements des actifs financiers. Elle constate que les prix des actifs financiers tendent à évoluer de manière synchrone par industrie sectorielle (par exemple, l'énergie, la technologie, la chimie, le divertissement, les télécommunications). Au sein d'un même groupe, par exemple celui des télécommunications, les cours des actions de deux entreprises peuvent croître et décroître de manière synchrone (corrélation des rendements), mais avec des caractéristiques très différentes, par exemple des sauts brusques pour l'un et des variations quasi-continues pour l'autre (distribution des rendements). Nous pouvons alors vouloir les distinguer.

Nous proposons dans cette thèse une distance simple qui travaille à partir de deux vecteurs représentant les deux séries temporelles à comparer et encodant l'information de corrélation et l'information de distribution [68]. Il se trouve que cette distance peut s'interpréter comme une distance en corrélation de Spearman (corrélation linéaire sur les rangs) et en distance d'Hellinger (distance classique entre densités). La motivation et la construction de cette distance s'appuie sur la théorie des copules et le théorème de Sklar.

Des distances entre copules pour une nouvelle mesure de dépendance

Nous avons proposé, dans la partie précédente, une représentation alternative pour les séries temporelles financières, mais la distance reste encore relativement élémentaire : en pratique, pour la partie "dépendance" de l'information, elle calcule une simple corrélation de rang (Spearman). Cette corrélation de rang (tout comme celle de Kendall) est plus robuste aux valeurs aberrantes que la version linéaire, et est également invariante par transformations strictement monotones des variables, propriété héritée de sa définition en fonction de la copule. Cependant, les corrélations de Spearman et Kendall peuvent être vues comme des projections simplistes de la copule sous-jacente. Nous proposons de définir des mesures de dépendance alternatives qui se fondent sur la comparaison de la copule empirique avec des copules bien choisies [136, 141]. Ainsi, nous pouvons imiter le comportement du coefficient de Spearman en prenant comme copules de référence les copules encodant la dépendance positive, la dépendance négative (les bornes de Fréchet-Hoeffding), et l'indépendance. En considérant une bonne distance entre copules (nous avons suggéré le transport optimal [144]), nous pouvons mesurer où se trouve la copule empirique sur la géodésique allant de l'indépendance à l'une des copules cibles (dépendance positive ou négative). Nous avons illustré comment ce nouveau coefficient de dépendance se comporte par rapport à son concurrent direct (de par la motivation de la construction) le coefficient de Spearman, mais aussi par rapport à l'état de l'art des coefficients de dépendance provenant de la littérature statistique mathématique et apprentissage automatique. L'approche étant très générale, il est possible de changer les copules références pour définir d'autres types de coefficients (par exemple, de dépendance en queue de distribution), ou encore adopter un point de vue plus "reconnaissance de forme" et chercher dans un jeu de données des couples de variables vérifiant un certain motif de dépendance encodé dans une copule cible.

Considérations pratiques

Enfin, nous discutons de quelques considérations pratiques, importantes pour pouvoir appliquer le clustering, telles que choisir un bon nombre de clusters, utiliser des outils de visualisations pour inspecter les clusters et leurs évolutions, ainsi que comparer différents résultats [142].

Bibliography

[1] Karim T Abou-Moustafa and Frank P Ferrie. "A Note on Metric Properties for Some Divergence Measures: The Gaussian Case." In: *ACML*. 2012, pp. 1–15.

[2] Martial Agueh and Guillaume Carlier. "Barycenters in the Wasserstein space." In: *SIAM Journal on Mathematical Analysis* 43.2 (2011), pp. 904–924.

[3] Shun-ichi Amari. "Divergence function, information monotonicity and information geometry." In: *WITMSE*. 2009.

[4] Sio Iong Ao et al. "CLUSTAG: hierarchical clustering and graph methods for selecting tag SNPs." In: *Bioinformatics* 21.8 (2005), pp. 1735–1736.

[5] David Arthur and Sergei Vassilvitskii. "k-means++: the advantages of careful seeding." In: *SODA '07: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. New Orleans, Louisiana: Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035. ISBN: 978-0-898716-24-5.

[6] Colin Atkinson and Ann FS Mitchell. "Rao's distance measure." In: *Sankhyā: The Indian Journal of Statistics, Series A* (1981), pp. 345–365.

[7] Eduard Baitinger and Jochen Papenbrock. "Interconnectedness Risk and Active Portfolio Management." In: (2016).

[8] Eduard Baitinger and Jochen Papenbrock. "Interconnectedness Risk and Active Portfolio Management: The Information-theoretic Perspective." In: (2017).

[9] Sivaraman Balakrishnan et al. "Noise thresholds for spectral clustering." In: *Advances in Neural Information Processing Systems*. 2011, pp. 954–962.

[10] Arindam Banerjee et al. "Clustering with Bregman divergences." In: *The Journal of Machine Learning Research* 6 (2005), pp. 1705–1749.

[11] Frédéric Barbaresco. "Geometric radar processing based on Fréchet distance: information geometry versus optimal transport theory." In: *2011 12th International Radar Symposium (IRS)*. 2011.

[12] Nicolas Basalto et al. "Hausdorff clustering of financial time series." In: *Physica A: Statistical Mechanics and its Applications* 379.2 (2007), pp. 635–644.

[13] H Ben Ameur, D Brigo, and E Errais. *Pricing credit default swaps Bermudan options: An approximate dynamic programming approach*. Tech. rep. Working paper, 2005.

[14] Shai Ben-David, Ulrike Von Luxburg, and Dávid Pál. "A sober look at clustering stability." In: *Learning theory*. Springer, 2006, pp. 5–19.

[15] Donald J Berndt and James Clifford. "Using Dynamic Time Warping to Find Patterns in Time Series." In: *KDD workshop*. Vol. 10. 16. Seattle, WA. 1994, pp. 359–370.

[16] Johan GB Beumee et al. "Charting a course through the CDS Big Bang." In: (2009).

[17] Jacob Bien and Robert Tibshirani. "Hierarchical clustering with prototypes via minimax linkage." In: *Journal of the American Statistical Association* 106.495 (2011), pp. 1075–1084.

[18] Monica Billio et al. “Econometric measures of connectedness and systemic risk in the finance and insurance sectors.” In: *Journal of Financial Economics* 104.3 (2012), pp. 535–559.

[19] Mikolaj Binkowski, Gautier Marti, and Philippe Donnat. “Autoregressive Convolutional Neural Networks for Asynchronous Time Series.” In: *arXiv preprint arXiv:1703.04122* (2017).

[20] G Bonanno, F Lillo, RN Mantegna, et al. “High-frequency cross-correlation in a set of stocks.” In: *Quantitative Finance* 1.1 (2001), pp. 96–104.

[21] Giovanni Bonanno et al. “Topology of correlation-based minimal spanning trees in real and model markets.” In: *Physical Review E* 68.4 (2003), p. 046130.

[22] Christian Borghesi, Matteo Marsili, and Salvatore Miccichè. “Emergence of time-horizon invariant correlation structure in financial returns by subtraction of the market mode.” In: *Physical Review E* 76.2 (2007), p. 026104.

[23] Petro Borysov, Jan Hannig, and JS Marron. “Asymptotics of hierarchical clustering for growing dimension.” In: *Journal of Multivariate Analysis* 124 (2014), pp. 465–479.

[24] Mike Bostock. “D3. js-Data-Driven Documents (2016).” In: *URL: https://d3js. org* (2016).

[25] Eike Christian Brechmann. “Hierarchical Kendall copulas: Properties and inference.” In: *Canadian Journal of Statistics* 42.1 (2014), pp. 78–108.

[26] Eike Christian Brechmann et al. “Hierarchical Kendall copulas and the modeling of systemic and operational risk.” PhD thesis. Universitätsbibliothek der TU München, 2013.

[27] Leo Breiman and Jerome H Friedman. “Estimating optimal transformations for multiple regression and correlation.” In: *Journal of the American statistical Association* 80.391 (1985), pp. 580–598.

[28] J Gabriel Brida and W Adrián Risso. “Hierarchical structure of the German stock market.” In: *Expert Systems with Applications* 37.5 (2010), pp. 3846–3852.

[29] Juan Gabriel Brida and Wiston Adrián Risso. “Multidimensional minimal spanning tree: The Dow Jones case.” In: *Physica A: Statistical Mechanics and its Applications* 387.21 (2008), pp. 5205–5210.

[30] D Brigo. “Constant Maturity CDS valuation with market models (2006).” In: *Risk Magazine, June issue. Earlier extended version available at http://ssrn. com/abstract* 639022 ().

[31] D Brigo. *Market models for CDS options and callable floaters*. 2005.

[32] D Brigo, A Pallavicini, and R Torresetti. “CDO calibration with the dynamical Generalized Poisson Loss model. ssrn. com.” In: *Published later in Risk Magazine* (2007).

[33] Damiano Brigo. “CDS options through candidate market models and the CDS-calibrated CIR++ stochastic intensity model.” In: *Credit Risk: Models, Derivatives and Management, Taylor & Francis* (2008).

[34] Damiano Brigo and Aurélien Alfonsi. “Credit default swap calibration and derivatives pricing with the SSRD stochastic intensity model.” In: *Finance and stochastics* 9.1 (2005), pp. 29–42.

[35] Damiano Brigo and Agostino Capponi. “Bilateral counterparty risk with application to CDSs.” In: *Risk* 23.3 (2010), p. 85.

[36] Damiano Brigo, Agostino Capponi, and Andrea Pallavicini. “Arbitrage-Free Bilateral Counterparty Risk Valuation Under Collateralization and Application to Credit Default Swaps.” In: *Mathematical Finance* 24.1 (2014), pp. 125–146.

[37] Damiano Brigo and Kyriakos Chourdakis. "Counterparty risk for credit default swaps: Impact of spread volatility and default correlation." In: *International Journal of Theoretical and Applied Finance* 12.07 (2009), pp. 1007–1026.

[38] Damiano Brigo and Naoufel El-Bachir. "An exact formula for default swaptionsŠ pricing in the SSRJD stochastic intensity model." In: *Mathematical Finance* 20.3 (2010), pp. 365–382.

[39] Damiano Brigo and Fabio Mercurio. *Interest rate models-theory and practice: with smile, inflation and credit*. Springer Science & Business Media, 2007.

[40] Damiano Brigo and Massimo Morini. "Structural credit calibration." In: *RISK-LONDON-RISK MAGAZINE LIMITED-* 19.4 (2006), p. 78.

[41] Damiano Brigo, Massimo Morini, and Andrea Pallavicini. *Counterparty credit risk, collateral and funding: with pricing cases for all asset classes*. John Wiley & Sons, 2013.

[42] Damiano Brigo and Andrea Pallavicini. "Counterparty risk and contingent CDS valuation under correlation between interest-rates and default." In: (2007).

[43] Damiano Brigo, Andrea Pallavicini, and Roberto Torresetti. "Cluster-based extension of the generalized Poisson loss dynamics and consistency with single names." In: *International Journal of Theoretical and Applied Finance* 10.04 (2007), pp. 607–631.

[44] Damiano Brigo, Andrea Pallavicini, and Roberto Torresetti. *Credit models and the crisis: A journey into CDOs, copulas, correlations and dynamic models*. John Wiley & Sons, 2010.

[45] Damiano Brigo, Andrea Pallavicini, and Roberto Torresetti. "Credit models and the crisis: An overview." In: *Journal of Risk Management in Financial Institutions* 4.3 (2011), pp. 243–253.

[46] Damiano Brigo and Marco Tarenghi. "Credit Default Swap Calibration with a Tractable Structural Model." In: *Proceedings of the FEA (Financial Engineering and Applications) 2004 Conference at MIT, Cambridge, Massachusetts, November*, pp. 8–10.

[47] Damiano Brigo et al. "Liquidity modeling for credit default swaps: An overview." In: *Credit Risk Frontiers: Subprime Crisis, Pricing and Hedging, CVA, MBS, Ratings, and Liquidity* (2011), pp. 585–617.

[48] Joël Bun et al. "Rotational invariant estimator for general noisy matrices." In: *IEEE Transactions on Information Theory* 62.12 (2016), pp. 7475–7490.

[49] Gunnar Carlsson and Facundo Mémoli. "Characterization, stability and convergence of hierarchical clustering methods." In: *Journal of machine learning research* 11.Apr (2010), pp. 1425–1470.

[50] Yale Chang et al. "A Robust-Equitable Copula Dependence Measure for Feature Selection." In: *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*. 2016, pp. 84–92.

[51] Zhenmin Chen and John W Van Ness. "Space-conserving agglomerative algorithms." In: *Journal of classification* 13.1 (1996), pp. 157–168.

[52] Moorad Choudhry. *An introduction to credit derivatives*. Butterworth-Heinemann, 2012.

[53] Moorad Choudhry. *The credit default swap basis*. Vol. 45. Bloomberg press New York, NY, 2006.

[54] Rama Cont. "Empirical properties of asset returns: stylized facts and statistical issues." In: (2001).

[55] Sueli IR Costa, Sandra A Santos, and João E Strapasson. "Fisher information distance: a geometrical reading." In: *Discrete Applied Mathematics* 197 (2015), pp. 59–69.

[56] Christopher L Culp, Andria van der Merwe, and Bettina J Staerkle. "Single-Name Credit Default Swaps: A Review of the Empirical Academic Literature." In: (2016).

[57] Chester Curme et al. "Emergence of statistically validated financial intraday lead-lag relationships." In: *Quantitative Finance* 15.8 (2015), pp. 1375–1386.

[58] Chester Curme et al. "How Lead-Lag Correlations Affect the Intraday Pattern of Collective Stock Dynamics." In: *Office of Financial Research Working Paper* 15-15 (2015).

[59] Marco Cuturi. "Sinkhorn distances: Lightspeed computation of optimal transport." In: *Advances in Neural Information Processing Systems*. 2013, pp. 2292–2300.

[60] Marco Cuturi and Arnaud Doucet. "Fast Computation of Wasserstein Barycenters." In: *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*. 2014, pp. 685–693. URL: http://jmlr.org/proceedings/papers/v32/cuturi14.html.

[61] Paul Deheuvels. "An asymptotic decomposition for multivariate distribution-free tests of independence." In: *Journal of Multivariate Analysis* 11.1 (1981), pp. 102–113.

[62] Paul Deheuvels. "La fonction de dépendance empirique et ses propriétés. Un test non paramétrique d'indépendance." In: *Acad. Roy. Belg. Bull. Cl. Sci.(5)* 65.6 (1979), pp. 274–292.

[63] T. Di Matteo, F. Pozzi, and T. Aste. "The use of dynamical networks to detect the hierarchical organization of financial market sectors." In: *The European Physical Journal B - Condensed Matter and Complex Systems* (Aug. 2010). ISSN: 1434-6028.

[64] Tiziana Di Matteo, Francesca Pozzi, and Tomaso Aste. "The use of dynamical networks to detect the hierarchical organization of financial market sectors." In: *The European Physical Journal B* 73.1 (2010), pp. 3–11.

[65] A Adam Ding and Yi Li. "Copula Correlation: An Equitable Dependence Measure and Extension of Pearson's Correlation." In: *arXiv preprint arXiv:1312.7214* (2013).

[66] Hui Ding et al. "Querying and Mining of Time Series Data: Experimental Comparison of Representations and Distance Measures." In: (2008).

[67] Philippe Donnat, Gautier Marti, and Philippe Very. "Toward a generic representation of random variables for machine learning." In: *Pattern Recognition Letters* 70 (2016), pp. 24–31. DOI: 10.1016/j.patrec.2015.11.004. URL: http://dx.doi.org/10.1016/j.patrec.2015.11.004.

[68] Philippe Donnat, Gautier Marti, and Philippe Very. "Toward a generic representation of random variables for machine learning." In: *Pattern Recognition Letters* 70 (2016), pp. 24–31.

[69] Christian Dose and Silvano Cincotti. "Clustering of financial time series with application to index and enhanced index tracking portfolio." In: *Physica A: Statistical Mechanics and its Applications* 355.1 (2005), pp. 145–151.

[70] S Drożdż et al. "Dynamics of competition between collectivity and noise in the stock market." In: *Physica A: Statistical Mechanics and its Applications* 287.3 (2000), pp. 440–449.

[71] Fabrizio Durante and Roberta Pappada. "Cluster analysis of time series via Kendall distribution." In: *Strengthening Links Between Data Analysis and Soft Computing*. Springer, 2015, pp. 209–216.

[72] Fabrizio Durante, Roberta Pappadà, and Nicola Torelli. "Clustering of financial time series in risky scenarios." In: *Advances in Data Analysis and Classification* 8.4 (2014), pp. 359–376.

[73] Fabrizio Durante, Susanne Saminger-Platz, and Peter Sarkoci. "Rectangular patchwork for bivariate copulas and tail dependence." In: *Communications in Statistics-Theory and Methods* 38.15 (2009), pp. 2515–2527.

[74] Fabrizio Durante et al. "A portfolio diversification strategy via tail dependence measures." In: (2015).

[75] Bradley Efron. "Bootstrap methods: another look at the jackknife." In: *Breakthroughs in Statistics*. Springer, 1992, pp. 569–593.

[76] Ahmed Drissi El Maliani et al. "Color texture classification using rao distance between multivariate copula based models." In: *Computer Analysis of Images and Patterns*. Springer. 2011, pp. 498–505.

[77] Gal Elidan. "Copulas in machine learning." In: *Copulae in Mathematical and Quantitative Finance*. Springer, 2013, pp. 39–60.

[78] Edwin J Elton and Martin J Gruber. "Improved forecasting through the design of homogeneous groups." In: *The Journal of Business* 44.4 (1971), pp. 432–450.

[79] Thomas W Epps. "Comovements in stock prices in the very short run." In: *Journal of the American Statistical Association* 74.366a (1979), pp. 291–298.

[80] Eugene F. Fama. "The Behavior of Stock-Market Prices." In: *The Journal of Business* 38.1 (1965), pp. 34–105. ISSN: 00219398.

[81] Paweł Fiedor. "Information-theoretic approach to lead-lag effect on financial markets." In: *The European Physical Journal B* 87.8 (2014), pp. 1–9.

[82] Paweł Fiedor. "Networks in financial markets based on the mutual information rate." In: *Physical Review E* 89.5 (2014), p. 052801.

[83] Gregory A Fredricks and Roger B Nelsen. "On the relationship between Spearman's rho and Kendall's tau for pairs of continuous random variables." In: *Journal of Statistical Planning and Inference* 137.7 (2007), pp. 2143–2150.

[84] Brendan J Frey and Delbert Dueck. "Clustering by passing messages between data points." In: *science* 315.5814 (2007), pp. 972–976.

[85] Tal Galili. "dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering." In: *Bioinformatics* (2015), btv428.

[86] Ya-Chun Gao, Yong Zeng, and Shi-Min Cai. "Influence network in the Chinese stock market." In: *Journal of Statistical Mechanics: Theory and Experiment* 2015.3 (2015), P03017.

[87] Christian Genest, JJ Molina, and JA Lallena. "De l'impossibilité de construire des lois à marges multidimensionnelles données à partir de copules." In: *Comptes rendus de l'Académie des sciences. Série 1, Mathématique* 320.6 (1995), pp. 723–726.

[88] Zoubin Ghahramani, Barnabás Póczos, and Jeff G Schneider. "Copula-based Kernel Dependency Measures." In: *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*. 2012, pp. 775–782.

[89] Lorenzo Giada and Matteo Marsili. "Algorithms of maximum likelihood data clustering with applications." In: *Physica A: Statistical Mechanics and its Applications* 315.3 (2002), pp. 650–664.

[90] Lorenzo Giada and Matteo Marsili. "Data clustering and noise undressing of correlation matrices." In: *Physical Review E* 63.6 (2001), p. 061101.

[91] Arthur Gretton et al. “A kernel two-sample test.” In: *Journal of Machine Learning Research* 13.Mar (2012), pp. 723–773.

[92] Mark A Hall. “Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning.” In: *Proceedings of the 17th International Conference on Machine Learning (ICML-00)*. 2000, pp. 359–366.

[93] Fang Han and Han Liu. “Optimal rates of convergence for latent generalized correlation matrix estimation in transelliptical distribution.” In: *arXiv preprint arXiv:1305.6916* (2013).

[94] Dion Harmon et al. “Networks of economic market interdependence and systemic risk.” In: *arXiv preprint arXiv:1011.3707* (2010).

[95] John A Hartigan. “Consistency of single linkage for high-density clusters.” In: *Journal of the American Statistical Association* 76.374 (1981), pp. 388–394.

[96] Keith Henderson, Brian Gallagher, and Tina Eliassi-Rad. “EP-MEANS: An Efficient Nonparametric Clustering of Empirical Probability Distributions.” In: (2015).

[97] Wenbo Hu and Alec N Kercheval. “Portfolio optimization for student t and skewed t returns.” In: *Quantitative Finance* 10.1 (2010), pp. 91–105.

[98] Feixue Huang, Pengfei Gao, and Yu Wang. “Comparison of Prim and Kruskal on Shanghai and Shenzhen 300 Index hierarchical structure tree.” In: *Web Information Systems and Mining, 2009. WISM 2009. International Conference on*. IEEE. 2009, pp. 237–241.

[99] Wei-Qiang Huang et al. “A financial network perspective of financial institutions’ systemic risk contributions.” In: *Physica A: Statistical Mechanics and its Applications* 456 (2016), pp. 183–196.

[100] L. Hubert and P. Arabie. “Comparing partitions.” In: *Journal of classification* 2.1 (1985), pp. 193–218.

[101] John C Hull. *Options, futures, and other derivatives*. Pearson Education, 2006.

[102] Neil F Johnson et al. “What shakes the FX tree? Understanding currency dominance, dependence, and dynamics (Keynote Address).” In: *SPIE Third International Symposium on Fluctuations and Noise*. International Society for Optics and Photonics. 2005, pp. 86–99.

[103] Woo-Sung Jung et al. “Group dynamics of the Japanese market.” In: *Physica A: Statistical Mechanics and its Applications* 387.2 (2008), pp. 537–542.

[104] Ioannis Katsavounidis, C-C Jay Kuo, and Zhen Zhang. “A new initialization technique for generalized Lloyd iteration.” In: *Signal Processing Letters, IEEE* 1.10 (1994), pp. 144–146.

[105] Bryan Kelly and Hao Jiang. “Tail risk and asset prices.” In: *Review of Financial Studies* 27.10 (2014), pp. 2841–2871.

[106] Dror Y Kenett et al. “Dependency network and node influence: application to the study of financial markets.” In: *International Journal of Bifurcation and Chaos* 22.07 (2012), p. 1250181.

[107] Dror Y Kenett et al. “Dominating clasp of the financial sector revealed by partial correlation analysis of the stock market.” In: *PloS one* 5.12 (2010), e15032.

[108] Dror Y Kenett et al. “Dynamics of stock market correlations.” In: *AUCO Czech Economic Review* 4.3 (2010), pp. 330–341.

[109] Azadeh Khaleghi et al. “Online clustering of processes.” In: *International Conference on Artificial Intelligence and Statistics*. 2012, pp. 601–609.

[110] Azadeh Khaleghi et al. “Online Clustering of Processes.” In: JMLR Proceedings 22 (2012). Ed. by Neil D. Lawrence and Mark Girolami, pp. 601–609.

[111] HJ Kim et al. “Scale-free network in stock markets.” In: *Journal-Korean Physical Society* 40 (2002), pp. 1105–1108.

[112] Benjamin F King. “Market and industry factors in stock price behavior.” In: *The Journal of Business* 39.1 (1966), pp. 139–190.

[113] Justin B Kinney and Gurinder S Atwal. “Equitability, mutual information, and the maximal information coefficient.” In: *Proceedings of the National Academy of Sciences* 111.9 (2014), pp. 3354–3359.

[114] Jon Kleinberg. “An impossibility theorem for clustering.” In: *Advances in neural information processing systems* (2003), pp. 463–470.

[115] Anton Kocheturov, Mikhail Batsyn, and Panos M Pardalos. “Dynamics of cluster structures in a financial market network.” In: *Physica A: Statistical Mechanics and its Applications* 413 (2014), pp. 523–533.

[116] Ravi Kothari and Dax Pitts. “On finding the number of clusters.” In: *Pattern Recognition Letters* 20.4 (1999), pp. 405–416.

[117] Akshay Krishnamurthy et al. “Efficient active algorithms for hierarchical clustering.” In: *International Conference on Machine Learning* (2012).

[118] L Kullmann, J Kertesz, and RN Mantegna. “Identification of clusters of companies in stock indices via Potts super-paramagnetic transitions.” In: *Physica A: Statistical Mechanics and its Applications* 287.3 (2000), pp. 412–419.

[119] Tilman Lange et al. “Stability-based validation of clustering solutions.” In: *Neural computation* 16.6 (2004), pp. 1299–1323.

[120] Delphine Lautier and Franck Raynaud. “Systemic risk in energy derivative markets: a graph-theory analysis.” In: *Available at SSRN 1579629* (2011).

[121] Gan Siew Lee and Maman A Djauhari. “Multidimensional stock network analysis: An Escoufier’s RV coefficient approach.” In: *AIP Conference Proceedings*. Vol. 1. 1557. 2013, pp. 550–555.

[122] Junghoon Lee, Janghyuk Youn, and Woojin Chang. “Intraday volatility and network topological properties in the Korean stock market.” In: *Physica A: Statistical mechanics and its Applications* 391.4 (2012), pp. 1354–1360.

[123] Victoria Lemieux et al. “Clustering techniques and their effect on portfolio formation and risk analysis.” In: *Proceedings of the International Workshop on Data Science for Macro-Modeling*. ACM. 2014, pp. 1–6.

[124] Erel Levine and Eytan Domany. “Resampling method for unsupervised estimation of cluster validity.” In: *Neural computation* 13.11 (2001), pp. 2573–2593.

[125] Haijun Li, Marco Scarsini, and Moshe Shaked. “Dynamic linkages for multivariate distributions with given nonoverlapping multivariate marginals.” In: *Journal of multivariate analysis* 68.1 (1999), pp. 54–77.

[126] Haijun Li, Marco Scarsini, and Moshe Shaked. “Linkages: a tool for the construction of multivariate distributions with given nonoverlapping multivariate marginals.” In: *Journal of Multivariate Analysis* 56.1 (1996), pp. 20–41.

[127] Eckhard Liebscher et al. “Copula-based dependence measures.” In: *Dependence Modeling* 2.1 (2014), pp. 49–64.

[128] Innar Liiv. “Seriation and matrix reordering methods: An historical overview.” In: *Statistical analysis and data mining* 3.2 (2010), pp. 70–91.

[129] Jessica Lin et al. "A symbolic representation of time series, with implications for streaming algorithms." In: *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*. ACM. 2003, pp. 2–11.

[130] Han Liu, Fang Han, and Cun-hui Zhang. "Transelliptical graphical models." In: *Advances in Neural Information Processing Systems*. 2012, pp. 809–817.

[131] Han Liu et al. "High-dimensional semiparametric Gaussian copula graphical models." In: *The Annals of Statistics* 40.4 (2012), pp. 2293–2326.

[132] David Lopez-Paz, Philipp Hennig, and Bernhard Schölkopf. "The Randomized Dependence Coefficient." In: *NIPS* (2013).

[133] Rosario N Mantegna. "Hierarchical structure in financial markets." In: *The European Physical Journal B-Condensed Matter and Complex Systems* 11.1 (1999), pp. 193–197.

[134] Rosario N Mantegna and H Eugene Stanley. *Introduction to econophysics: correlations and complexity in finance*. Cambridge university press, 1999.

[135] Martin Martens and Ser-Huang Poon. "Returns synchronization and daily correlation dynamics between international stock markets." In: *Journal of Banking & Finance* 25.10 (2001), pp. 1805–1827.

[136] Gautier Marti, Frank Nielsen, and Philippe Donnat. "Optimal copula transport for clustering multivariate time series." In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2016, pp. 2379–2383.

[137] Gautier Marti et al. "A Proposal of a Methodological Framework with Experimental Guidelines to Investigate Clustering Stability on Financial Time Series." In: *14th IEEE International Conference on Machine Learning and Applications, ICMLA 2015, Miami, FL, USA, December 9-11, 2015*. 2015, pp. 32–37. DOI: 10.1109/ICMLA.2015.11. URL: http://dx.doi.org/10.1109/ICMLA.2015.11.

[138] Gautier Marti et al. "A review of two decades of correlations, hierarchies, networks and clustering in financial markets." In: *arXiv preprint arXiv:1703.00485* (2017).

[139] Gautier Marti et al. "Clustering Financial Time Series: How Long Is Enough?" In: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*. 2016, pp. 2583–2589. URL: http://www.ijcai.org/Abstract/16/367.

[140] Gautier Marti et al. "Clustering Random Walk Time Series." In: *International Conference on Networked Geometric Science of Information*. Springer. 2015, pp. 675–684.

[141] Gautier Marti et al. "Exploring and measuring non-linear correlations: Copulas, Lightspeed Transportation and Clustering." In: *Proceedings of the NIPS 2016 Time Series Workshop, co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016*. 2016, pp. 59–69. URL: http://jmlr.org/proceedings/papers/v55/marti16.html.

[142] Gautier Marti et al. "HCMapper: An interactive visualization tool to compare partition-based flat clustering extracted from pairs of dendrograms." In: *arXiv preprint arXiv:1507.08137* (2015).

[143] Gautier Marti et al. "On clustering financial time series: a need for distances between dependent random variables." In: *Computational Information Geometry*. Springer, 2017, pp. 149–174.

[144] Gautier Marti et al. "Optimal transport vs. Fisher-Rao distance between copulas for clustering multivariate time series." In: *IEEE Statistical Signal Processing Workshop, SSP 2016, Palma de Mallorca, Spain, June 26-29, 2016*. 2016, pp. 1–5. DOI: 10.1109/SSP.2016.7551770. URL: http://dx.doi.org/10.1109/SSP.2016.7551770.

[145] David Matesanz and Guillermo J Ortega. "Sovereign public debt crisis in Europe. A network analysis." In: *Physica A: Statistical Mechanics and its Applications* 436 (2015), pp. 756–766.

[146] Sergio Mayordomo, Juan Ignacio Peña, and Eduardo S Schwartz. "Are all credit default swap databases equal?" In: *European Financial Management* 20.4 (2014), pp. 677–713.

[147] Nicolai Meinshausen and Peter Bühlmann. "Stability selection." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72.4 (2010), pp. 417–473.

[148] Hao Meng et al. "Systemic risk and spatiotemporal dynamics of the US housing market." In: *Scientific Reports* 4 (2014), p. 3655.

[149] Glenn W Milligan and Martha C Cooper. "An examination of procedures for determining the number of clusters in a data set." In: *Psychometrika* 50.2 (1985), pp. 159–179.

[150] Gaspard Monge. *Mémoire sur la théorie des déblais et des remblais*. De l'Imprimerie Royale, 1781.

[151] Stefano Monti et al. "Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data." In: *Machine learning* 52.1-2 (2003), pp. 91–118.

[152] Raffaello Morales, T Di Matteo, and Tomaso Aste. "Dependency structure and scaling properties of financial time series are related." In: *Scientific Reports* 4.4589 (2014).

[153] Fionn Murtagh and Pedro Contreras. "Algorithms for hierarchical clustering: an overview." In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2.1 (2012), pp. 86–97.

[154] Nicoló Musmeci, Tomaso Aste, and T Di Matteo. "Risk diversification: a study of persistence with a filtered correlation-network approach." In: *Network Theory in Finance* 1.1 (2015), pp. 77–98.

[155] Nicolo Musmeci, Tomaso Aste, and Tiziana Di Matteo. "Relation between Financial Market Structure and the Real Economy: Comparison between Clustering Methods." In: *Available at SSRN 2525291* (2014).

[156] Roger B Nelsen. *An introduction to copulas*. Vol. 139. Springer Science & Business Media, 2013.

[157] Frank Nielsen and Frederic Barbaresco. *Geometric Science of Information: Second International Conference, GSI 2015, Palaiseau, France, October 28-30, 2015, Proceedings*. Vol. 9389. Springer, 2015.

[158] Frank Nielsen and Richard Nock. "On approximating the smallest enclosing Bregman balls." In: *Proceedings of the twenty-second annual symposium on Computational geometry*. ACM. 2006, pp. 485–486.

[159] Dominic O'Kane. *Modelling single-name and multi-name credit derivatives*. Vol. 573. John Wiley & Sons, 2011.

[160] J-P Onnela, Kimmo Kaski, and Janos Kertész. "Clustering and information in correlation based financial networks." In: *The European Physical Journal B-Condensed Matter and Complex Systems* 38.2 (2004), pp. 353–362.

[161] J-P Onnela et al. "Dynamic asset trees and Black Monday." In: *Physica A: Statistical Mechanics and its Applications* 324.1 (2003), pp. 247–252.

[162] J-P Onnela et al. "Dynamic asset trees and portfolio analysis." In: *The European Physical Journal B-Condensed Matter and Complex Systems* 30.3 (2002), pp. 285–288.

[163] J-P Onnela et al. "Dynamics of market correlations: Taxonomy and portfolio analysis." In: *Physical Review E* 68.5 (2003), p. 056110.

[164] Raj Kumar Pan and Sitabhra Sinha. "Collective behavior of stock price movements in an emerging market." In: *Physical Review E* 76.4 (2007), p. 046116.

[165] Don B Panton, V Parker Lessig, and O Maurice Joy. "Comovement of international equity markets: a taxonomic approach." In: *Journal of Financial and Quantitative Analysis* 11.03 (1976), pp. 415–432.

[166] Jochen Papenbrock and Peter Schwendner. "Handling risk-on/risk-off dynamics with correlation regimes and correlation networks." In: *Financial Markets and Portfolio Management* 29.2 (2015), pp. 125–147.

[167] Fabian Pedregosa et al. "Scikit-learn: Machine learning in Python." In: *Journal of Machine Learning Research* 12.Oct (2011), pp. 2825–2830.

[168] Xavier Pennec. "Statistical computing on manifolds: from Riemannian geometry to computational anatomy." In: *Emerging Trends in Visual Computing*. Springer, 2009, pp. 347–386.

[169] Gustavo Peralta and Abalfazl Zareei. "A network approach to portfolio selection." In: *Journal of Empirical Finance* (2016).

[170] Donald B Percival and Andrew T Walden. *Wavelet methods for time series analysis*. Vol. 4. Cambridge University Press, 2006.

[171] Vasiliki Plerou et al. "A random matrix theory approach to financial cross-correlations." In: *Physica A: Statistical Mechanics and its Applications* 287.3 (2000), pp. 374–382.

[172] Barnabás Póczos, Zoubin Ghahramani, and Jeff Schneider. "Copula-based Kernel Dependency Measures." In: (2012).

[173] David Pollard et al. "Strong consistency of k-means clustering." In: *The Annals of Statistics* 9.1 (1981), pp. 135–140.

[174] Francesco Pozzi, Tiziana Di Matteo, and Tomaso Aste. "Spread of risk across financial markets: better to invest in the peripheries." In: *Scientific reports* 3 (2013).

[175] BCR Quantitative Credit Strategy. "CDS Curve Trading Handbook 2008." In: *Barclays Capital Research* (2008).

[176] C Radhakrishna Rao. "Information and the accuracy attainable in the estimation of statistical parameters." In: *Breakthroughs in statistics*. Springer, 1992, pp. 235–247.

[177] Fei Ren et al. "Dynamic portfolio strategy using clustering approach." In: *arXiv preprint arXiv:1608.03058* (2016).

[178] David Reshef et al. "Equitability analysis of the maximal information coefficient, with comparisons." In: *arXiv preprint arXiv:1301.6314* (2013).

[179] David N Reshef et al. "Detecting novel associations in large data sets." In: *science* 334.6062 (2011), pp. 1518–1524.

[180] Jacopo Rocchi, Enoch Yan Lok Tsui, and David Saad. "Emerging interdependence between stock values during financial crashes." In: *arXiv preprint arXiv:1611.02549* (2016).

[181] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. "The earth mover's distance as a metric for image retrieval." In: *International journal of computer vision* 40.2 (2000), pp. 99–121.

[182] D. Ryabko. "Clustering processes." In: *Proc. the 27th International Conference on Machine Learning (ICML 2010)*. Haifa, Israel, 2010, pp. 919–926.

[183] D. Ryabko. "Clustering processes." In: *Proc. the 27th International Conference on Machine Learning (ICML 2010)*. Haifa, Israel, 2010, pp. 919–926.

[184] Daniil Ryabko. “Clustering processes.” In: (2010).

[185] Stan Salvador and Philip Chan. “Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms.” In: *Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference on*. IEEE. 2004, pp. 576–584.

[186] Romeil Sandhu, Tryphon Georgiou, and Allen Tannenbaum. “Market fragility, systemic risk, and Ricci curvature.” In: *arXiv preprint arXiv:1505.05182* (2015).

[187] Dino Sejdinovic et al. “Equivalence of distance-based and RKHS-based statistics in hypothesis testing.” In: *The Annals of Statistics* 41.5 (2013), pp. 2263–2291.

[188] Ahmet Sensoy and Benjamin M Tabak. “Dynamic spanning trees in stock market networks: The case of Asia-Pacific.” In: *Physica A: Statistical Mechanics and its Applications* 414 (2014), pp. 387–402.

[189] Ohad Shamir and Naftali Tishby. “Cluster Stability for Finite Samples.” In: *NIPS*. 2007.

[190] Ohad Shamir and Naftali Tishby. “Model selection and stability in k-means clustering.” In: (2008).

[191] Noah Simon and Robert Tibshirani. “Comment on" Detecting Novel Associations In Large Data Sets" by Reshef Et Al, Science Dec 16, 2011.” In: *arXiv preprint arXiv:1401.7645* (2014).

[192] A. Sklar. “Fonctions de répartition à n dimensions et leurs marges.” In: (1959).

[193] A Sklar. *Fonctions de répartition à n dimensions et leurs marges*. Université Paris 8, 1959.

[194] Reginald Smith. “The spread of the credit crisis: view from a stock correlation network.” In: *Journal of the Korean Physical Society* 54.6 (2009), pp. 2460–2463.

[195] Dong-Ming Song et al. “Evolution of worldwide stock markets, correlation structure, and correlation-based graphs.” In: *Physical Review E* 84.2 (2011), p. 026108.

[196] Won-Min Song, T Di Matteo, and Tomaso Aste. “Hierarchical information clustering by means of topologically embedded graphs.” In: *PLoS One* 7.3 (2012), e31929.

[197] Won-Min Song, Di Matteo, and Tomaso Aste. “Hierarchical Information Clustering by Means of Topologically Embedded Graphs.” In: *PLoS ONE* 7.3 (Mar. 2012), e31929+.

[198] Bharath K Sriperumbudur et al. “Kernel Choice and Classifiability for RKHS Embeddings of Probability Distributions.” In: *NIPS*. 2009, pp. 1750–1758.

[199] Catherine A Sugar and Gareth M James. “Finding the number of clusters in a dataset: An information-theoretic approach.” In: *Journal of the American Statistical Association* 98.463 (2003), pp. 750–763.

[200] Marc Szafraniec, Gautier Marti, and Philippe Donnat. “Putting Self-Supervised Token Embedding on the Tables.” In: *arXiv preprint arXiv:1708.04120* (2017).

[201] Gábor J Székely, Maria L Rizzo, et al. “Brownian distance covariance.” In: *The annals of applied statistics* 3.4 (2009), pp. 1236–1265.

[202] Asuka Takatsu et al. “Wasserstein geometry of gaussian measures.” In: *Osaka Journal of Mathematics* 48.4 (2011), pp. 1005–1026.

[203] Hideki Takayasu. *Practical fruits of econophysics*. Springer, 2006.

[204] Yoshikazu Terada. “Strong consistency of factorial K-means clustering.” In: *Annals of the Institute of Statistical Mathematics* 67.2 (2013), pp. 335–357.

[205] Yoshikazu Terada. “Strong Consistency of Reduced K-means Clustering.” In: *Scandinavian Journal of Statistics* 41.4 (2014), pp. 913–931.

[206] Robert Tibshirani, Guenther Walther, and Trevor Hastie. "Estimating the number of clusters in a data set via the gap statistic." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.2 (2001), pp. 411–423.

[207] Vincenzo Tola et al. "Cluster analysis for portfolio optimization." In: *Journal of Economic Dynamics and Control* 32.1 (2008), pp. 235–258.

[208] Roberto Torresetti, Damiano Brigo, and Andrea Pallavicini. "Risk-neutral versus objective loss distribution and CDO tranche valuation." In: *Journal of Risk Management in Financial Institutions* 2.2 (2009), pp. 175–192.

[209] Bence Tóth and János Kertész. "Accurate estimator of correlations between asynchronous signals." In: *Physica A: Statistical Mechanics and its Applications* 388.8 (2009), pp. 1696–1705.

[210] Joel A Tropp. "An introduction to matrix concentration inequalities." In: *arXiv preprint arXiv:1501.01571* (2015).

[211] Chengyi Tu. "Cointegration-based financial networks study in Chinese stock market." In: *Physica A: Statistical Mechanics and its Applications* 402 (2014), pp. 245–254.

[212] M. Tumminello, F. Lillo, and Mantegna. "Correlation, hierarchies, and networks in financial markets." In: *Journal of Economic Behaviour and Organization* (2010).

[213] M Tumminello, RN Mantegna, and F Lillo. "Shrinkage and Spectral Filtering of Correlation Matrices: A Comparison via the Kullback-Leibler Distance." In: *Acta Physica Polonica. Series B* 39.1 (2008), pp. 4079–4088.

[214] M. Tumminello et al. "A tool for filtering information in complex systems." In: *Proceedings of the National Academy of Sciences USA* 102 (2005), pp. 10421–10426.

[215] Michele Tumminello, Fabrizio Lillo, and Rosario N Mantegna. "Correlation, hierarchies, and networks in financial markets." In: *Journal of Economic Behavior & Organization* 75.1 (2010), pp. 40–58.

[216] Michele Tumminello, Fabrizio Lillo, and Rosario N Mantegna. "Hierarchically nested factor model from multivariate data." In: *EPL (Europhysics Letters)* 78.3 (2007), p. 30006.

[217] Michele Tumminello, Fabrizio Lillo, and Rosario N Mantegna. "Kullback-Leibler distance as a measure of the information filtered from multivariate data." In: *Physical Review E* 76.3 (2007), p. 031123.

[218] Michele Tumminello et al. "A tool for filtering information in complex systems." In: *Proceedings of the National Academy of Sciences of the United States of America* 102.30 (2005), pp. 10421–10426.

[219] Michele Tumminello et al. "Correlation based networks of equity returns sampled at different time horizons." In: *The European Physical Journal B* 55.2 (2007), pp. 209–217.

[220] Michele Tumminello et al. "Spanning trees and bootstrap reliability estimation in correlation-based networks." In: *International Journal of Bifurcation and Chaos* 17.07 (2007), pp. 2319–2329.

[221] Michele Tumminello et al. "Statistically validated networks in bipartite complex systems." In: *PloS one* 6.3 (2011), e17994.

[222] N Vandewalle, F Brisbois, X Tordoir, et al. "Non-random topology of stock markets." In: *Quantitative Finance* 1.3 (2001), pp. 372–374.

[223] Cédric Villani. *Optimal transport: old and new*. Vol. 338. Springer Science & Business Media, 2008.

[224] Ulrike Von Luxburg, Mikhail Belkin, and Olivier Bousquet. "Consistency of spectral clustering." In: *The Annals of Statistics* (2008), pp. 555–586.

[225] Tomáš Vỳrost, Štefan Lyócsa, and Eduard Baumöhl. "Granger causality stock market networks: Temporal proximity and preferential attachment." In: *Physica A: Statistical Mechanics and its Applications* 427 (2015), pp. 262–276.

[226] Joe H. Ward. "Hierarchical Grouping to Optimize an Objective Function." In: *Journal of the American Statistical Association* 58.301 (1963), pp. 236–244.

[227] Richard White. "The pricing and risk management of credit default swaps, with a focus on the isda model." In: *OpenGamma Quantitative Research* 16 (2013).

[228] Lei Yu and Huan Liu. "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution." In: *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*. 2003, pp. 856–863.

[229] Xin Zhang et al. "Systemic risk and causality dynamics of the world international shipping market." In: *Physica A: Statistical Mechanics and its Applications* 415 (2014), pp. 43–53.

[230] Yiting Zhang et al. "Will the US economy recover in 2010? A minimal spanning tree study." In: *Physica A: Statistical Mechanics and its Applications* 390.11 (2011), pp. 2020–2050.

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : Quelques contributions aux méthodes de partitionnement automatique des séries temporelles financières, et applications aux couvertures de défaillance

Mots clefs : Partitionnement automatique, séries temporelles financières, copules, mesures de corrélations, distances entre distributions, stabilité des grappes

Résumé :
Nous commençons cette thèse par passer en revue l'ensemble épars de la littérature sur les méthodes de partitionnement automatique des séries temporelles financières. Ensuite, tout en introduisant les jeux de données qui ont aussi bien servi lors des études empiriques que motivé les choix de modélisation, nous essayons de donner des informations intéressantes sur l'état du marché des couvertures de défaillance peu connu du grand public sinon pour son rôle lors de la crise financière mondiale de 2007-2008. Contrairement à la majorité de la littérature sur les méthodes de partitionnement automatique des séries temporelles financières, notre but n'est pas de décrire et expliquer les résultats par des explications économiques, mais de pouvoir bâtir des modèles et autres larges systèmes d'information sur ces groupes homogènes. Pour ce faire, les fondations doivent être stables. C'est pourquoi l'essentiel des travaux entrepris et décrits dans cette thèse visent à affermir le bien-fondé de l'utilisation de ces regroupements automatiques en discutant de leur consistance et stabilité aux perturbations. De nouvelles distances entre séries temporelles financières prenant mieux en compte leur nature stochastique et pouvant être mis à profit dans les méthodes de partitionnement automatique existantes sont proposées. Nous étudions empiriquement leur impact sur les résultats. Les résultats de ces études peuvent être consultés sur `www.datagrapple.com`.

Title : Some contributions to the clustering of financial time series and applications to credit default swaps

Keywords : clustering, financial time series, copulas, measures of correlations, distances between distributions, stability of clusters

Abstract :
In this thesis we first review the scattered literature about clustering financial time series. We then try to give as much colors as possible on the credit default swap market, a relatively unknown market from the general public but for its role in the contagion of bank failures during the global financial crisis of 2007-2008, while introducing the datasets that have been used in the empirical studies. Unlike the existing body of literature which mostly offers descriptive studies, we aim at building models and large information systems based on clusters which are seen as basic building blocks: These foundations must be stable. That is why the work undertaken and described in the following intends to ground further the clustering methodologies. For that purpose, we discuss their consistency and propose alternative measures of similarity that can be plugged in the clustering methodologies. We study empirically their impact on the clusters. Results of the empirical studies can be explored at `www.datagrapple.com`.

Ecole Polytechnique
Route de Saclay, 91128 Palaiseau, France