Tracing the temporal map of clusters in the stock market

Argimiro Arratia (UPC), Alejandra Cabaña (UAB)
argimiro@lsi.upc.edu, acabana@mat.uab.cat

Coloquio, USB, Caracas, 25 Nov. 2011

Paper available at arXiv http://arxiv.org/abs/1111.3127

Goal

Identifying stocks whose return history strongly resembles each other in order to design investment strategies.

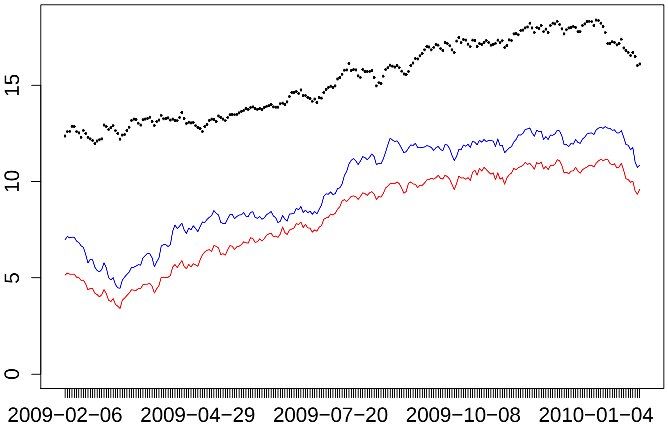

Figure: TEF, BBVA, SAN

More specific

- correlation depends on the length of time span
- interest in short time correlations, sustained through long periods of time
- Hence, we seek to capture the *correlation in movement*

Our model

A Temporal Graph representation of the evolution of clusters through time = a weighted graph whose vertices are clusters of stocks, bound by a relation of similarity on their return history, and edges among clusters with non empty intersection are weighted by the cardinality of this intersection.

Combinatorics on the TG of clusters

We can apply classical combinatorics of graphs to the TGC, with implications to investment strategies.

- Find paths of heaviest weight in the graph, translates to detecting the most stable clusters through time. In the context of finance it helps to detect those stocks whose return history mirror each other through long stretches of time.
- Finding a vertex cover, translates into finding a (minimum) set of stocks that intersects all clusters in the temporal graph. This set of stocks can be considered as a representative for the market index; or, it can be considered as a minimum portfolio covering the market.

Model design

For each stock X, we observe n values, $\mathbf{x} = (x_1, \ldots, x_n)$ corresponding to the *returns* during n trading days: $x_k = \dfrac{P(k)}{P(k-1)} - 1$, where $P(k)$ is the closing price of the stock at time k.

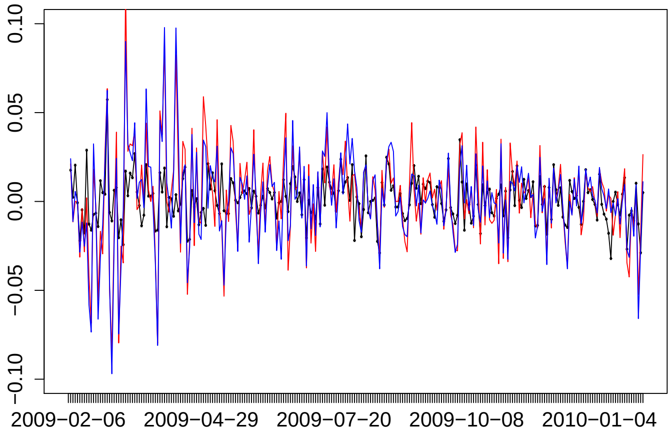

Figure: TEF, BBVA, SAN log returns

Compute the sample correlation coefficients for pairs of stocks returns, taken in a common period of n time instants:

$$r(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=1}^{n}(x_i - \bar{\mathbf{x}})(y_i - \bar{\mathbf{y}})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{\mathbf{x}})^2 \sum_{i=1}^{n}(y_i - \bar{\mathbf{y}})^2}}$$

where $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ are, respectively, the mean values of the observed samples.

$r = r(\mathbf{x},\mathbf{y})$ is a consistent estimator of the true correlation $\rho(X,Y)$

Under normal distribution, $\rho(X, Y) = 0$ imply X and Y independent; hence is useful to test the *null hypothesis* $H_0 : \rho = 0$.
We apply the test

$$t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

which is distributed as a t with $n-2$ degrees of freedom under the null hypothesis $\rho = 0$. Thus, "big" values of t_0 (or r) will lead us to reject the null hypothesis.
Table 1 shows the critical points c_α for the tests with critical region $|r| > c_\alpha$ for testing H_0, for levels $\alpha = 0.05$ and 0.01 for different values of n, that is, the number of trading days considered.

n	5	10	20	40	60	240
$c_{0.05}$	0.8783	0.6319	0.4438	0.3120	0.2542	0.1267
$c_{0.01}$	0.9587	0.7646	0.5614	0.4026	0.3301	0.1660

Table: Critical points for the contrast for $H_0 : \rho = 0$ with critical region $\{|r| > c_\alpha\}$

Clustering

We need a measure of proximity for a collection of returns

Warning:

- correlation values do not constitute a metric
- there is a bias implicit in the order in which return series are selected

We use correlation as a relation of similarity; build neighbourhoods of similar stocks and define a distance metric which amounts to saying that *two stocks are near, not only if they are correlated but also correlated to almost all similar stocks*

Clustering

Let τ be a time period.

- If one wants to cluster stocks with positive correlated returns, taken in the span of τ, then for each stock A, with series of returns **a** in the time period τ, we group into $\mathcal{G}_A$ all stocks X whose series of returns **x**, on the same time span of **a**, have correlation with **a** higher than a positive δ; i.e.,

  $$X \in \mathcal{G}_A \iff r(\mathbf{a}, \mathbf{x}) > \delta.$$

  δ is determined by our statistical test (cf. Table 1) and depends on the sample size. For example, for 40 days period we take $\delta = 0.65$, a midpoint between 1 and the threshold of $c_{0.05} = 0.3120$, to ensure some significant correlation.
  For the case of clustering stocks with negative correlated return time series, consider a negative δ and define $\mathcal{G}_A$ as $X \in \mathcal{G}_A \iff r(\mathbf{a}, \mathbf{x}) < \delta$.

- Next, for each pair of stocks A and B define

  $$d(A, B) = 1 - \frac{|\mathcal{G}_A \cap \mathcal{G}_B|}{|\mathcal{G}_A \cup \mathcal{G}_B|}$$

- Apply the *hierarchical clustering* algorithm to the collection of stocks, with respect to the distance metric d. Note that d depends on n, the number of trading sessions.

Repeat computations for further temporal intervals, and build a (weighted) representation of the evolution of clusters in time.

Temporal map of clusters

$\mathcal{S} = \{$ market stocks (time series) $\}$.
T the time period, partitioned into m successive sub-periods of time $\tau_1, \tau_2, \ldots, \tau_m$.
For $i = 1, \ldots, m$, let $\mathcal{C}_i$ be collection of clusters obtained in τ_i,
Let $S_{i,j}$, $1 \le i \le m$, $1 \le j \le |\mathcal{C}_i|$, be the clusters in $\mathcal{C}_i$ with at least two elements, and Q_i be the subset of elements with no significant correlation in time segment τ_i, for $1 \le i \le m$.
We define a directed graph $\mathcal{G}$ with vertex set the collection of subsets

$$\{S_{i,j} : 1 \le i \le m, 1 \le j \le |\mathcal{C}_i|\} \cup \{Q_i : 1 \le i \le m\}$$

and weighted edge set

$$\{(S_{i,j}, S_{i+1,k}, |S_{i,j} \cap S_{i+1,k}|) : 1 \le i \le m-1, 1 \le j \le |\mathcal{C}_i|, 1 \le k \le |\mathcal{C}_{i+1}|\}$$

We call $\mathcal{G}$ the *Temporal Graph of Clusters* (TGC) for the set of stocks $\mathcal{S}$ in the time segmentation $T = \bigcup_{1 \le i \le m} \tau_i$.

Computational results

$\mathcal{S} =$ IBEX35 $= \{$ ABE, ABG, ACS, ACX, ANA, BKT, BBVA , BME, BTO, CRI, ELE, ENG, FCC, FER, GAM, GAS, GRF, IBE, IBLA, IBR, IDR, ITX, MAP, POP, REE, REP, SAB, SAN, MTS, SYV, TEF, TL5, TRE, OHL $\}$

$T =$ 2008/06–2009/08

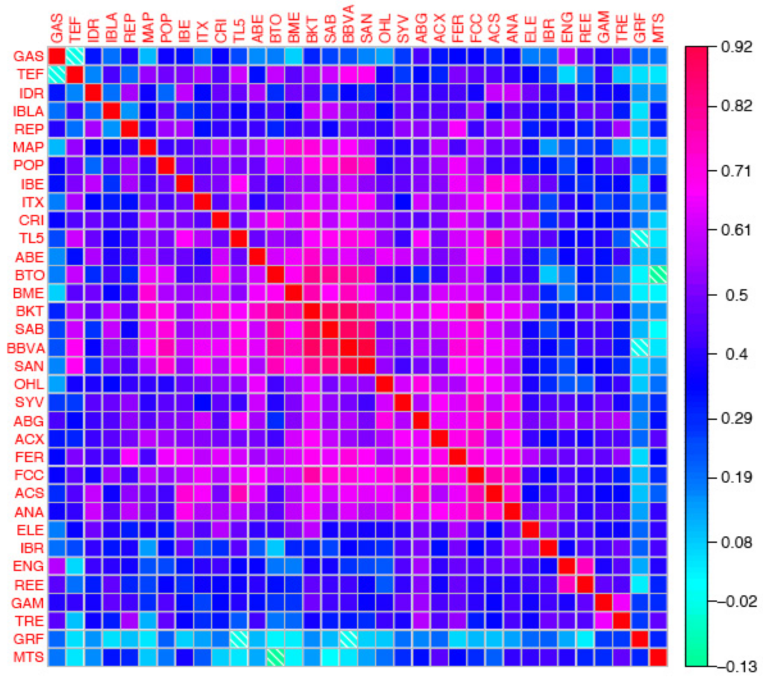

Figure: Correlation matrix for time segment 1-6-2008 to 1-8-2008

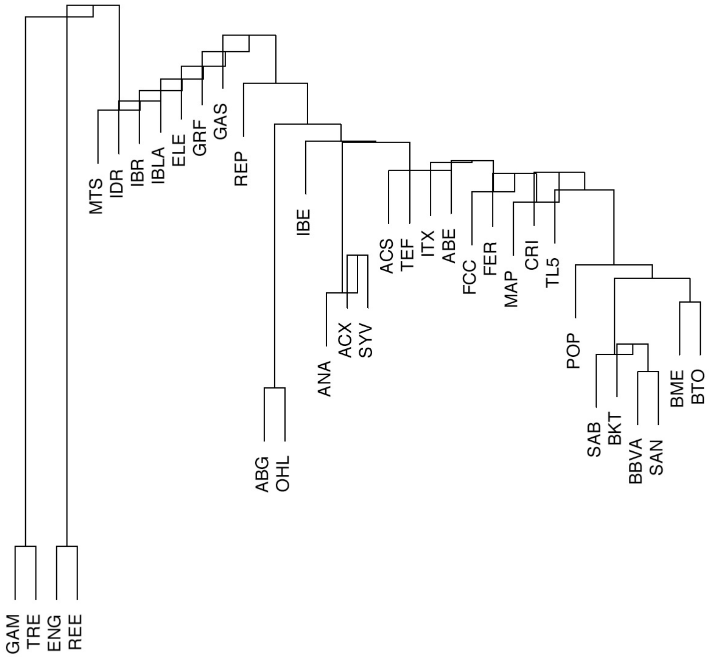

Figure: Dendogram for the correlation matrix

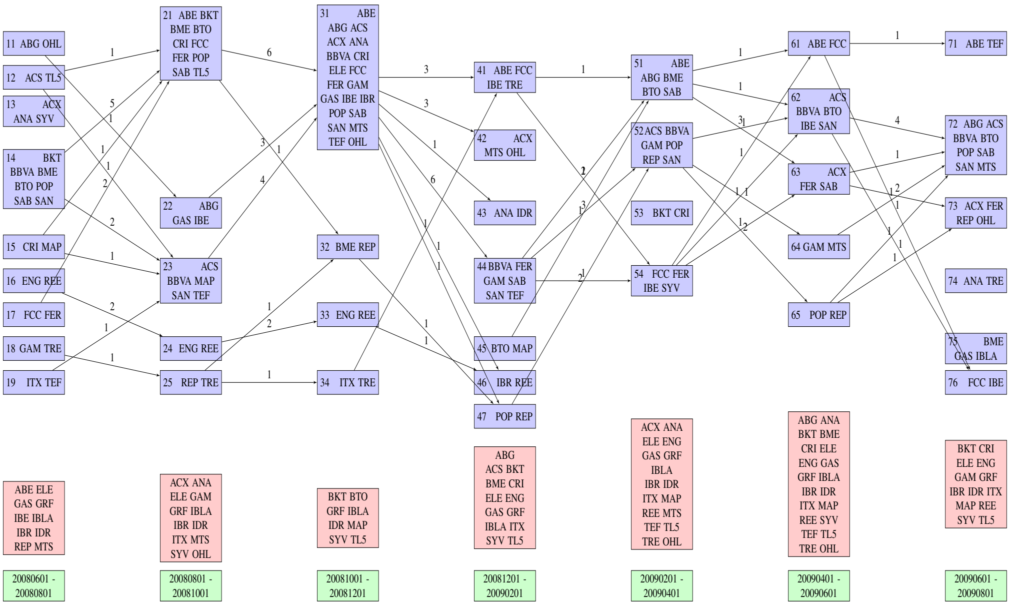
11 ABG OHL
12 ACS TL5
13 ACX ANA SYV
14 BKT BBVA BME BTO POP SAB SAN
15 CRI MAP
16 ENG REE
17 FCC FER
18 GAM TRE
19 ITX TEF
ABE ELE GAS GRF IBE IBLA IBR IDR REP MTS
20080601 - 20080801
21 ABE BKT BME BTO CRI FCC FER POP SAB TL5
22 ABG GAS IBE
23 ACS BBVA MAP SAN TEF
24 ENG REE
25 REP TRE
ACX ANA ELE GAM GRF IBLA IBR IDR ITX MTS SYV OHL
20080801 - 20081001
31 ABE ABG ACS ACX ANA BBVA CRI ELE FCC FER GAM GAS IBE IBR POP SAB SAN MTS TEF OHL
32 BME REP
33 ENG REE
34 ITX TRE
BKT BTO GRF IBLA IDR MAP SYV TL5
20081001 - 20081201
41 ABE FCC IBE TRE
42 ACX MTS OHL
43 ANA IDR
44 BBVA FER GAM SAB SAN TEF
45 BTO MAP
46 IBR REE
47 POP REP
ABG ACS BKT BME CRI ELE ENG GAS GRF IBLA ITX SYV TL5
20081201 - 20090201
51 ABE ABG BME BTO SAB
52 ACS BBVA GAM POP REP SAN
53 BKT CRI
54 FCC FER IBE SYV
ACX ANA ELE ENG GAS GRF IBLA IBR IDR ITX MAP REE MTS TEF TL5 TRE OHL
20090201 - 20090401
61 ABE FCC
62 ACS BBVA BTO IBE SAN
63 ACX FER SAB
64 GAM MTS
65 POP REP
ABG ANA BKT BME CRI ELE ENG GAS GRF IBLA IBR IDR ITX MAP REE SYV TEF TL5 TRE OHL
20090401 - 20090601
71 ABE TEF
72 ABG ACS BBVA BTO POP SAB SAN MTS
73 ACX FER REP OHL
74 ANA TRE
75 BME GAS IBLA
76 FCC IBE
BKT CRI ELE ENG GAM GRF IBR IDR ITX MAP REE SYV TL5
20090601 - 20090801

Figure: Temporal Graph of Clusters for IBEX 35, years 2008-2009

Combinatorics

- Stable clusters (Computing k-heaviest paths): Given a positive integer $k \geq 1$ and TGC $\mathcal{G} = \langle V, E \rangle$, with $V = \{S_{i,j} : 1 \leq i \leq m, 1 \leq j \leq n_i\}$, we want to find the k first heaviest paths starting at any cluster $S_{i,j}$ with no ingoing edge.
- Stock cover (Computing Vertex Cover): Given a TGC $\mathcal{G} = \langle V, E \rangle$, with $V = \{S_{i,j} : 1 \leq i \leq m, 1 \leq j \leq n_i\}$, we want to find a (minimum) set of stocks that intersects every (non singleton) cluster in $\mathcal{G}$.

Stable clusters

Let $\{\tau_1, \ldots, \tau_m\}$ be the sequence of time segments. Add a sink, i.e., an extra (dummy) cluster $S_{m+1,1}$ after time period τ_m.
Compute paths . . .

Brute force solution (bad)

Find all paths that end in $S_{m+1,1}$ (say by Breadth First Search), order them by weight and keep the k heaviest.
This incur in exponentially many comparisons: Worst case contains $O(n^m)$ many paths, where n is the maximum number of clusters per time segment and m is the number of time segments.

Dynamic programming solution (good)

Step i: at time segment τ_i define, for each $S_{i,j}$, a heap where the root, labelled $S_{i,j}$, has key 0, its descendent will be all heaps defined in previous time segments $\tau_{i'}$, $i' < i$, for which there is an edge from $S_{i',j'}$ to $S_{i,j}$, and the keys of all *leaves* of these descendent heaps are updated by adding the weight of the edge $(S_{i',j'}, S_{i,j})$.
Hence, the total weight of a path in the heap of root $S_{i,j}$ is the key of the leaf or initial node of such path. Order the leaves by key value and keep the paths for the k leaves with highest key, removing the paths of the remaining leaves before proceeding to step $i + 1$.
Running time: $O(mn^3)$.

Stock cover

The problem is NP-hard, so we apply a

Greedy strategy

Pick the stock that shows up more often in the collection of clusters V; then remove the clusters containing this stock (i.e. covered by the stock). Repeat with remaining clusters.

- To know which stock shows up more often, build an incidence matrix $\mathcal{M}$, where rows are labelled by the companies (stocks) and columns labelled by the clusters.
- Running time: $O(dim(\mathcal{M})) = O(|\mathcal{S}| \cdot |V|)$, where $\mathcal{S}$ is the set of stocks.
- The size of the solution set is at most one half of the size of the set of companies $\mathcal{S}$

Experimental results

Stable clusters

$k = 3$ *first heaviest paths* for IBEX during 2008-2009:

$$14 \overset{5}{\mapsto} 21 \overset{6}{\mapsto} 31 \overset{6}{\mapsto} 44 \overset{3}{\mapsto} 52 \overset{3}{\mapsto} 62 \overset{4}{\mapsto} 72$$

$$14 \overset{2}{\mapsto} 23 \overset{4}{\mapsto} 31 \overset{6}{\mapsto} 44 \overset{3}{\mapsto} 52 \overset{3}{\mapsto} 62 \overset{4}{\mapsto} 72$$

$$14 \overset{5}{\mapsto} 21 \overset{6}{\mapsto} 31 \overset{3}{\mapsto} 41 \overset{2}{\mapsto} 54 \overset{1}{\mapsto} 62 \overset{4}{\mapsto} 72$$

Stock cover

A stock cover for IBEX for the years 2008-2009:
{BBVA, FCC, REP, REE, TRE, ACX, ABG, CRI, TEF, MAP, BME, ANA, MTS, TL5}

Some conspicuous conclusions

- Santander (SAN) and BBVA, conform the most stable cluster through any period of time
- SAN and BBVA tend to participate in the largest clusters of IBEX35 components at different periods of time
- SAN, BBVA, TEF, REP and very few others move the Spanish market
- Negative correlation almost zero among IBEX35; hence an investor can not have a balanced portfolio consisting only on the big caps