Correlation-based Networks in Finance

Rosario Nunzio Mantegna
Palermo University, Italy

Observatory of Complex Systems
Work done in collaboration with

- Claudia Coronnello
- **Fabrizio Lillo**
- Salvatore Miccichè
- **Michele Tumminello**

- Tomaso Aste (ANU)
- Tiziana Di Matteo (ANU)
Overview

How to quantify and model information present in a correlation matrix?

- Describe the structure of an empirical correlation matrix

- Model hierarchies in terms of hierarchical trees and networks
Cross Correlation

N data series of length T

\[ r_i(t_j), \ j = 1, \ldots, T; \ i = 1, \ldots, N \]

Pearson’s correlation coefficient:

\[
\rho_{ij} = \frac{\langle r_i r_j \rangle - \langle r_i \rangle \langle r_j \rangle}{\sqrt{\langle r_i^2 \rangle - \langle r_i \rangle^2} \sqrt{\langle r_j^2 \rangle - \langle r_j \rangle^2}}
\]

Example:

Log-return of stock price

\[ r_i(t) \equiv \ln P_i(t) - \ln P_i(t - \tau) \]

Other correlation estimators:

- Fourier estimator
- ML correlation estimator
- Dynamical estimator
- …

Correlation Matrix

\[ C = \left( \rho_{ij} \right) \]
Statistical reliability of estimated cross correlation coefficients

\( N T \) data \( \rightarrow \sim N^2 \) correlation coefficients: a statistical uncertainty is associated with estimated correlation coefficients.

**Even in the best case we therefore need:**
- to *filter* (select) statistically reliable information

**Additional problems**

- Dynamics of correlations: non stationary process
- Interaction with environment.
- Heterogeneity of sampling
- Complexity of interactions
Complexity of correlation matrix

...by capitalization  ...by average linkage cluster analysis

\( N = 300 \) (NYSE), daily returns, 2001 - 2003, \( T = 748 \)
Clustering techniques and correlation based networks

**Clustering**  e.g. Hierarchical Clustering (Single linkage, Average linkage, etc)
Super Paramagnetic Clustering (E. Domany et al)
Maximum Likelihood Clustering (M. Marsili et al)
Sorting Point Into Neighbors (E. Domany et al)

**Correlation Based**  e.g. Minimum Spanning Tree (MST)
Networks
Average Linkage Minimum Spanning Tree
Planar Maximally Filtered Graph (PMFG)
Hierarchical clustering

A selection (computation) of \( n-1 \) correlation coefficient values from \( n(n-1)/2 \) distinct correlation coefficient values.

\[
C(s = 0) = C; \quad n - 1 \text{ construction steps}
\]

\[
C(s) \rightarrow \rho_{ij}(s) = \max_{h,k} [\rho_{hk}(s)] \rightarrow q(s + 1) = i(s) \cup j(s)
\]

Redefine the matrix \( C \) at the construction step \( s + 1 \) according to the rule \( R \)

\[
R(ALCA): \quad \rho_{qk} = \frac{N_i \rho_{ik} + N_j \rho_{jk}}{N_i + N_j}; \quad R(SLCA): \quad \rho_{qk} = \max(\rho_{ik}, \rho_{jk})
\]

Average linkage cluster analysis \quad Single linkage cluster analysis
Hierarchical clustering output

$N = 100$ (NYSE) daily returns 1995 - 1998

Average Linkage Cluster Analysis

$$C^< = (\rho_{ij}^<)$$

$$\rho_{ij}^< = \rho_{\alpha_k}$$

where $\alpha_k$ is the first node where elements $i$ and $j$ merge together
Filtered matrix

\[ N = 300 \text{ (NYSE); daily returns 2001 - 2003} \]

\[ C^< \text{ from average linkage cluster analysis} \]

\[ C \]
Another example (100 stocks of NYSE 95-98)

Single linkage cluster analysis correlation matrix

Sample correlation matrix
When one uses the stock order of the hierarchical tree the correlation matrix assumes a better readability.

The sample matrix is richer of information.
How to quantitatively compare correlation matrices obtained from different “filtering” methods?

We are interested to
(i) the amount of information retained by the filtering procedure;
(ii) its stability.
Filtered correlation matrices

We consider two filtered correlation matrices, obtained by applying the ALCA and the SLCA to the empirical correlation matrix respectively.

\[ C_{\text{ALCA}}^\prec \text{ and } C_{\text{SLCA}}^\prec \]

For comparison we also consider filtered correlation matrices obtained with Principal Component Analysis and Random Matrix Theory (RMT).
Kullback-Leibler distance

\[ K(p, q) = E_p \left[ \log \left( \frac{p}{q} \right) \right], \text{ where } p \text{ and } q \text{ are pdf's.} \]

Mutual information:

\[ I(X, Y) = K(p(X, Y), p(X)p(Y)) \]

For multivariate normal distributed random variables we have:

\[ K(P(\Sigma_1, X), P(\Sigma_2, X)) = \frac{1}{2} \left[ \log \left( \frac{|\Sigma_2|}{|\Sigma_1|} \right) + \text{tr} \left( \Sigma_2^{-1} \Sigma_1 \right) - n \right] = K(\Sigma_1, \Sigma_2) \]

Minimizing the Kullback-Leibler distance is equivalent to maximize the likelihood in the maximum likelihood factor analysis.

Expectation values

\[
E[K(\Sigma, S_1)] = \frac{1}{2} \left\{ n \log \left( \frac{2}{T} \right) + \sum_{p=T-n+1}^{T} \frac{\Gamma^I(p/2)}{\Gamma(p/2)} + \frac{n(n+1)}{T-n-1} \right\}
\]

\[
E[K(S_1, \Sigma)] = \frac{1}{2} \left\{ n \log \left( \frac{T}{2} \right) - \sum_{p=T-n+1}^{T} \frac{\Gamma^I(p/2)}{\Gamma(p/2)} \right\}
\]

\[
E[K(S_1, S_2)] = \frac{1}{2} \frac{n(n+1)}{T-n-1}
\]

where \( \Sigma \) is the true correlation matrix of the system while \( S_1 \) and \( S_2 \) are sample matrices of \( \Sigma \) from two independent realizations of length \( T \).

The Kullback-Leibler distance is a good estimator of statistical uncertainty (for normally distributed random variables).
Comparison of filtered correlation matrices

Increasing stability

\[ \langle k(C_{\text{filt}}(i), C_{\text{filt}}(j)) \rangle \]

- Block diagonal model with 12 factors.
- \(N=100,\) \(T=748.\)
- Gaussian random variables.
Comparison of filtered correlation matrices

- $N = 100$ (NYSE)
- Daily returns
- $T = 748$
Correlation based networks

\( (i, j, \rho_{ij}) \)

\[
C = \begin{pmatrix}
    1 & 0.13 & 0.90 & 0.81 \\
    0.13 & 1 & 0.57 & 0.34 \\
    0.90 & 0.57 & 1 & 0.71 \\
    0.81 & 0.34 & 0.71 & 1
\end{pmatrix}
\]

Correlation Matrix \((C)\)

\[
S = \begin{pmatrix}
    1 & 3 & 0.90 \\
    1 & 4 & 0.81 \\
    3 & 4 & 0.71 \\
    2 & 3 & 0.57 \\
    2 & 4 & 0.34 \\
    1 & 2 & 0.13
\end{pmatrix}
\]

Sorted List of Links \((S)\)
Minimum Spanning Tree

1. Define a similarity measure between the elements of the system

2. Construct the list \( S \) by ordering similarities in decreasing order

3. Starting from the first element of \( S \), add the corresponding link if and only if the graph is still a Forest or a Tree

Minimum Spanning Tree
MST
Minimum Spanning Tree (MST)

$N = 100$ (NYSE) daily returns 1995 - 1998

$T = 1011$

Define a similarity measure between the elements of the system

Construct the list $S$ by ordering similarities in decreasing order

Starting from the first element of $S$, add the corresponding link if and only if the graph is still a Forest or a Tree

Minimum Spanning Tree MST

Starting from the first element of $S$, add the corresponding link if and only if the graph is still Planar ($g=0$)

Planar Maximally Filtered Graph PMFG

Planar Maximally Filtered Graph (PMFG)

$N = 100$ (NYSE) daily returns 1995 - 1998

$T = 1011$
Cliques in the PMFG

When \( g = 0 \), topological constraints allow the observation of \textbf{cliques} of 3 and 4 vertices only.
Are the detected links statistically reliable?

\[ N = 300 \text{ (NYSE)} \]
\[ 2001 - 2003 \]
\[ T = 748 \]
A validation based on bootstrap

<table>
<thead>
<tr>
<th></th>
<th>Data Set</th>
<th>Pseudo-replicate Data Set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$e_1$</td>
<td>$e_2$</td>
</tr>
<tr>
<td>$t_1$</td>
<td>0.113</td>
<td>1.123</td>
</tr>
<tr>
<td>$t_2$</td>
<td>1.567</td>
<td>0.789</td>
</tr>
<tr>
<td>$t_3$</td>
<td>1.065</td>
<td>-1.962</td>
</tr>
<tr>
<td>$t_4$</td>
<td>1.112</td>
<td>0.998</td>
</tr>
<tr>
<td>$t_5$</td>
<td>-0.211</td>
<td>0.312</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$T$</td>
<td>0.479</td>
<td>-1.828</td>
</tr>
</tbody>
</table>

M surrogated data matrices are constructed, e.g. $M=1000$. 
Bootstrap value of nodes

bootstrap value distribution

ALCA

bootstrap value
Statistical reliability of the minimum spanning tree

$N = 300$ (NYSE) daily returns
2001 - 2003
$T = 748$

Planar Maximally Filtered Graph

NYSE
$N = 100$
daily
returns
2002
$T = 256$
Bootstrap vs correlation

For Gaussian series: $\sigma_\rho = \frac{1 - \rho^2}{\sqrt{T - 3}}$

$N = 300$ (NYSE)
daily returns
2001 - 2003
$T = 748$
Topology of correlation based financial networks

Topology† of MSTs in

- empirical data;
- and
- in a one-factor model.

†Bonanno, Caldarelli, Lillo and Mantegna, PRE 68, 046130 (2003).
In spite of the fact that the one-factor model explains more than 85% of the elements of correlation matrix.
Empirical data

1071 stocks continuously traded at the NYSE during the period 1987-1998 (3030 trading days)

The color code refers to the SIC index:
- finance
- manufacturing
- construction
- utilities
- wholesale trade
- mining
- retail trade
- services
- public administration
MST of a one-factor model

\[ R_i(t) = \alpha_i + \beta_i R_M(t) + \varepsilon_i(t) \]

However, the topology of the MST of the one-factor model is completely different from the one of real data.
Conclusions

We have investigated the structure of an empirical correlation matrix by using hierarchical trees and correlation based networks.

We have introduced the Kullback-Leibler distance in order to compare different techniques used to filter the most stable information of correlation matrices.

We have estimated the statistical reliability of links in hierarchical trees and correlation based networks by using a bootstrap based approach.

We have investigated correlation based networks in financial market. They are able to provide information about the collective behavior of different stocks and about models able to describe it.
OCS website: http://ocs.dft.unipa.it