Hellebore
Capital Management

On Clustering Financial Time Series

Gautier Marti, Philippe Donnat and Frank Nielsen

Noisy correlation matrices

Let X be the matrix storing the standardized returns of $N = 560$ assets (credit default swaps) over a period of $T = 2500$ trading days.

Then, the empirical correlation matrix of the returns is

$$C = \frac{1}{T} X X^\top.$$

We can compute the empirical density of its eigenvalues

$$\rho(\lambda) = \frac{1}{N} \frac{dn(\lambda)}{d\lambda},$$

where $n(\lambda)$ counts the number of eigenvalues of C less than λ.

From random matrix theory, the **Marchenko-Pastur distribution** gives the limit distribution as $N \to \infty, T \to \infty$ and T/N fixed. It reads:

$$\rho(\lambda) = \frac{T/N}{2\pi} \frac{\sqrt{(\lambda_{\max} - \lambda)(\lambda - \lambda_{\min})}}{\lambda},$$

where $\lambda_{\min}^{\max} = 1 + N/T \pm 2\sqrt{N/T}$, and $\lambda \in [\lambda_{\min}, \lambda_{\max}]$.

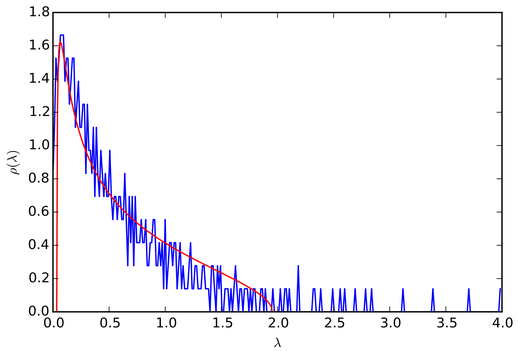

Figure 1: Marchenko-Pastur density vs. empirical density of the correlation matrix eigenvalues

Notice that the Marchenko-Pastur density fits well the empirical density meaning that most of the information contained in the empirical correlation matrix amounts to noise: only 26 eigenvalues are greater than $\lambda_{\max}$.
The highest eigenvalue corresponds to the 'market', the 25 others can be associated to 'industrial sectors'.

Clustering time series

Given a correlation matrix of the returns,

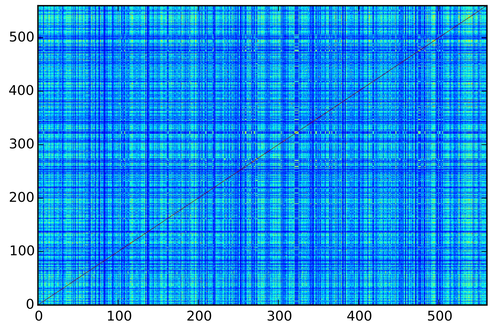

Figure 2: An empirical and noisy correlation matrix

one can re-order assets using a hierarchical clustering algorithm to make the hierarchical correlation pattern blatant,

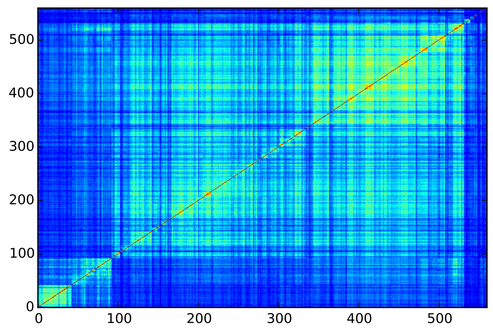

Figure 3: The same noisy correlation matrix re-ordered by a hierarchical clustering algorithm

and finally filter the noise according to the correlation pattern:

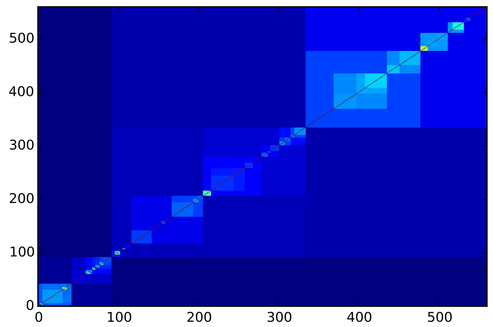

Figure 4: The resulting filtered correlation matrix

Beyond correlation

Sklar's Theorem. For any random vector $X = (X_1, \ldots, X_N)$ having continuous marginal cumulative distribution functions F_i, its joint cumulative distribution F is uniquely expressed as

$$F(X_1, \ldots, X_N) = C(F_1(X_1), \ldots, F_N(X_N)),$$

where C, the multivariate distribution of uniform marginals, is known as the copula of X.

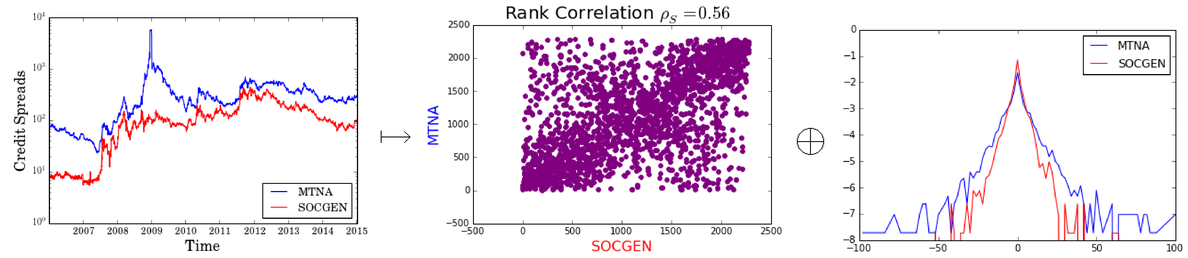

Figure 5: ArcelorMittal and Société générale prices are projected on dependence $\oplus$ distribution space; notice their heavy-tailed exponential distribution.

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 cdf. We define the following distance

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

where $d_1^2(G_X(X), G_Y(Y)) = 3\mathbb{E}[|G_X(X) - G_Y(Y)|^2]$, and $d_0^2(G_X, G_Y) = \frac{1}{2} \int_{\mathbf{R}} \left(\sqrt{\frac{dG_X}{d\lambda}} - \sqrt{\frac{dG_Y}{d\lambda}} \right)^2 d\lambda.$

Clustering Results & Stability

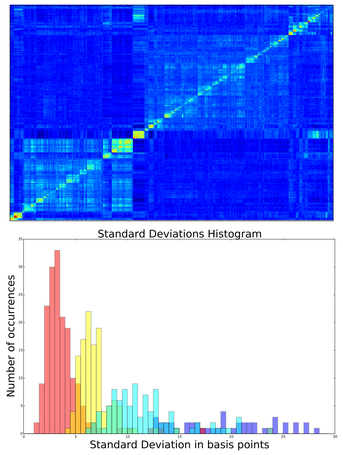

Figure 6: (Top) The returns correlation structure appears more clearly using rank correlation; (Bottom) Clusters of returns distributions can be partly described by the returns volatility

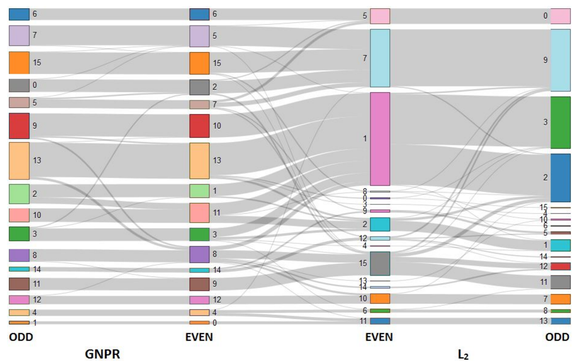

Figure 7: Stability test on Odd/Even trading days subsampling: our approach (GNPR) yields more stable clusters with respect to this perturbation than standard approaches (using Pearson correlation or L_2 distances).