Guided Tour of Machine Learning in Finance

Week 3: Unsupervised Learning

Clustering and estimation of equity correlation matrix

Igor Halperin

NYU Tandon School of Engineering, 2017

Estimation of equity correlation matrix

Let $\mathbf{x}_i = \left\{ x_i^{(t)} \right\}_{t=1}^{T}$, $i = 1,\ldots,N$ be a vector of standardized log-returns of stock .

Empirical correlation matrix: $\hat{C}_{ij} = \frac{1}{T-1} \sum_{t=1}^{T} x_i^{(t)} x_j^{(t)}$

Estimation of equity correlation matrix

Let $\mathbf{x}_i = \left\{ x_i^{(t)} \right\}_{t=1}^{T}$, $i = 1,\ldots,N$ be a vector of standardized log-returns of stock .

Empirical correlation matrix: $\hat{C}_{ij} = \frac{1}{T-1} \sum_{t=1}^{T} x_i^{(t)} x_j^{(t)}$

The number of free parameters : $N(N-1)/2 \sim N^2/2$

The number of observations: NT

Estimation of equity correlation matrix

Let $\mathbf{x}_i = \left\{ x_i^{(t)} \right\}_{t=1}^{T}$, $i = 1,\ldots,N$ be a vector of standardized log-returns of stock .

Empirical correlation matrix: $\hat{C}_{ij} = \frac{1}{T-1} \sum_{t=1}^{T} x_i^{(t)} x_j^{(t)}$

The number of free parameters : $N(N-1)/2 \sim N^2/2$
The number of observations: NT

should have $T \gg N$ for a reliable estimation

Estimation of equity correlation matrix

Let $\mathbf{x}_i = \left\{ x_i^{(t)} \right\}_{t=1}^{T}$, $i = 1,\ldots,N$ be a vector of standardized log-returns of stock .

Empirical correlation matrix: $\hat{C}_{ij} = \frac{1}{T-1} \sum_{t=1}^{T} x_i^{(t)} x_j^{(t)}$

The number of free parameters : $N(N-1)/2 \sim N^2/2$
The number of observations: NT

should have $T \gg N$ for a reliable estimation

The "true" correlation matrix may differ from the empirical correlation matrix due to the "observational noise", whose amount depends on the ratio T/N

Estimation of equity correlation matrix

Let $\mathbf{x}_i = \left\{ x_i^{(t)} \right\}_{t=1}^{T}$, $i = 1,\ldots,N$ be a vector of standardized log-returns of stock .

Empirical correlation matrix: $\hat{C}_{ij} = \frac{1}{T-1} \sum_{t=1}^{T} x_i^{(t)} x_j^{(t)}$

The number of free parameters : $N(N-1)/2 \sim N^2/2$
The number of observations: NT

should have $T \gg N$ for a reliable estimation

The "true" correlation matrix may differ from the empirical correlation matrix due to the "observational noise", whose amount depends on the ratio T/N

Estimation of a "noise-filtered" equity correlation matrix is important for:
- trading
- portfolio management
- systemic risk monitoring

Estimation of equity correlation matrix

Some de-noising methods:

1. Methods based on the Random Matrix Theory (RMT) (E. Stanley et. al. 1999)

2. “Shrinkage” methods (Ledoit and Wold 2003)

3. Clustering-based filtering: based on distances between individual points, and aggregation of linkages between sub-clusters

Estimation of equity correlation matrix

Clustering:
Given N items with pair-wise distances d_{ij}, divide them into K groups so that minimum distance between items in different groups is maximized.

Maximum minimum distance:
-maintain clusters as a set of connected components of a graph
- iteratively combine the clusters containing the two closest items by adding an edge between them
- stop when there are K clusters

Estimation of equity correlation matrix

Clustering:
Given N items with pair-wise distances d_{ij}, divide them into K groups so that minimum distance between items in different groups is maximized.

Maximum minimum distance:
-maintain clusters as a set of connected components of a graph
- iteratively combine the clusters containing the two closest items by adding an edge between them
- stop when there are K clusters

This is the **Kruskal algorithm**, an example of a **single-linkage agglomerative clustering.**

Estimation of equity correlation matrix

Clustering:
Given N items with pair-wise distances d_{ij}, divide them into K groups so that minimum distance between items in different groups is maximized.

Maximum minimum distance:
-maintain clusters as a set of connected components of a graph
- iteratively combine the clusters containing the two closest items by adding an edge between them
- stop when there are K clusters

This is the **Kruskal algorithm**, an example of a **single-linkage agglomerative clustering.**

Pair-wise distances: $$d_{ij} = \sqrt{2(1-\rho_{ij})}$$

This choice fulfills the three axioms of an Euclidean metric:
- $d_{ij} = 0 \; iff \; i = j$
- $d_{ij} = d_{ji}$
- $d_{ij} \leq d_{ik} + d_{kj}, \quad \forall i,j,k$

Control question

Select all correct answers

1. Correlation matrices are most reliably estimated when $T \ll N$.
2. A clustering-based filtering of correlation matrices is based on distances between individual points, and aggregation of linkages between sub-clusters
3. The Kruskal algorithm belongs in the class of Complete Linkage methods.
4. Pair-wise distances in clustering-based methods for correlation matrices are determined in terms of pair-wise correlations.

Correct answers: 2, 4