CSS07
Pattern recognition’s role in the articulation and validation of Complex Structures
Part 1: Unsupervised Learning

New York Stock Exchange Network

Known cyclic structure

Gregory Leibon
Dartmouth College
Goals:

• A basic introduction to pattern recognition
  • Dimension reduction, supervised learning, unsupervised learning, graphical models (See: Pattern Classification by Duda, Hart, and Stork and The Elements of Statistical Learning by Hastie, Tibshirani, Friedman)

• Some specific tools to use in the next couple week!
  • In MATLAB: MDS, K-means, HMM, Basic Kernel Methods
  • Others: Spectral Clustering, Persistent Homology, Bayesian Networks

• Explore a Concrete Example: Chunk of U.S. Equity’s Market
  • The market analysis presented here is a brief look into current work in progress with Alyssa Anderson, Robert Savell, Daniel Rockmore, and Scott Pauls.
Market as a Complex System

• Long Term Goal

• is to form a model of the market with simple local behavior out of which emerges the complex collective behavior of the market we see in the world.

• Role of Pattern Recognition

• is that in a realistic model one must understand how to articulate the local behavior as well understand the emergent patterns that characterize the a complex system.

• Data Collected

• 6000+ interacting tradable equities from NASDAQ and NYSE from Yahoo! finance.

• Data collected for each equity was the Open, Close, High, and Low price as well as the Volume traded on each trading day over the last 15 years.

• Annotation (partial) collected was the equity’s names, sector, industry, and index membership.
An Equity’s Times Series

\[ X_t = \frac{C_t - C_{t-1}}{C_{t-1}} \approx d(\ln(S_t)) \]

A well known rough approximation:

\[ d(\ln(S_t)) = \sigma(t)dB_t + c(t)dt \]
Our Market as a Network

- Bonds
- Currency
- MBS
- Commodities

Equities

NASDAQ
NYSE
AMEX
Correlation Metric

Normalize:

$$\hat{X} = \frac{X - \langle X \rangle}{\sqrt{\langle (X - \langle X \rangle)^2 \rangle}}$$

Correlation:

$$\rho(X, Y) = \langle \hat{X} \hat{Y} \rangle$$

Metric:

$$D(X, Y) = \arccos(\rho(X, Y))$$

Corresponds to a metric graph in $$S^{1000 \pm c}$$
Correlation: S&P500

How much of this correlation is internal to our equities market and how much due to external forces from the whole economy?
\[ d(\ln(S_t)) = \sigma dB_t + c dt + \tau dF_t \]
Confounding Factors

Each day there is pressure on the equities market to absorb money. We can remove this effect:
Network Structure

• First step: **Dimension Reduction**

• Our network is embedded in 1000+ dimensions. To see and work with our network, it is useful to attempt to embed it in a lower dimensional space.

• One great simple tool for doing this is **Multi Dimensional Scaling, MDS**.

• With MDS: For each dimension, we can produce an approximate lower dimensional embedding, call it \( f(X) \).

• Might use the Euclidean:  
\[
d'(X,Y) = \sin\left(\frac{d(X,Y)}{2}\right)
\]
MDS algorithm

Simply, attempt to minimize a positive loss function that would be zero if for all \( X, Y \)

\[
d(X,Y) = d(f(X), f(Y)).
\]

Example Raw Stress: \( L = \sum (d(X,Y) - d(f(X), f(Y)))^2 \)

Minimization Techniques: Gradient Decent, Newton Raphson, Iterative Majorization, Tabu Search, Genetic Algorithms, Simulated Annealing....
Himalayas
2-d Example

Example taken from MatLab's help
MDS 2-d

![MDS 2-d Diagram](image-url)
For S&P500

Call the embedding $f$ and use the Euclidean distance $d(f(X), f(Y))$

$$ S = \frac{\sum |d(X,Y) - d(f(X), f(Y))|}{\sum |d(X,Y)|} $$
MATLAB code:

Here Closes is the 1000 by 500 matrix with columns corresponding to each equity’s normalized daily Close.

\[
\text{Cor}=\text{corrcoef}(\text{Closes},'\text{rows}',\text{'pairwise'})
\]

\[
\text{Dist}=\sin(\text{acos}(\text{Cor}/2));
\]

\[
\text{opt}=\text{statset}('\text{MaxIter}',5000);
\]

\[
\text{Dim15}=\text{mdscale}(\text{Dist},15,'\text{Options}',\text{opt});
\]

\%
Here Y15 is a fifteen dimension embedding of vertices.
Network Shape

- Re-scale network
- Clustering to find nodes at new scale.
- Unsupervised learning.
- Clustering techniques: kmeans, spectral clustering, *-Linkage Clustering, Delaunay Complex Exploitation Algorithms,...
K-means

- Simplest clustering algorithm is \textit{k-means}
- To run requires fixing \( N = \#(\text{Clusters}) \)
- Requires an Euclidean type embedding (we have one via the MDS)
- Once again, we are essentially minimizing a loss function:

\[
L = \sum_{k=1}^{N} \sum_{x_i \in C_k} (x_i - \mu_k)^2
\]
K-means algorithm:

1. Randomly choose points in each cluster and compute centroids.

2. Organize points by distance to the centroids.
3. Update centroids
4. Repeat...
...until stable.
MATLAB

Recall Dim15 was the 15 dimensional Euclidean realization of correlation information.

Clusters = kmeans(Dim15,20);

Clusters is a list of 500 labels between 1 and 20, one for each point in Dim15, indicating the cluster that each sample has been associated with.
Hardest part is choosing $N = \#(\text{Clusters})$

Elbowlogy:

Elbow
**Laplacian**

\[ L = \begin{bmatrix}
1 & -1/2 & -1/2 & 0 & 0 & 0 & 0 & 0 & 0 \\
-1/2 & 1 & -1/2 & 0 & 0 & 0 & 0 & 0 & 0 \\
-1/2 & -1/2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & -1/4 & -1/4 & -1/4 & -1/4 & 0 \\
0 & 0 & 0 & -1/4 & 1 & -1/4 & -1/4 & -1/4 & 0 \\
0 & 0 & 0 & -1/4 & -1/4 & 1 & -1/4 & -1/4 & 0 \\
0 & 0 & 0 & -1/4 & -1/4 & -1/4 & 1 & -1/4 & 0 \\
0 & 0 & 0 & -1/4 & -1/4 & -1/4 & -1/4 & 1 & 0 \\
0 & 0 & 0 & -1/4 & -1/4 & -1/4 & -1/4 & -1/4 & 1
\end{bmatrix} \]
Laplacian

\[ d'(X,Y) = \sin \left( \frac{\rho(X,Y)}{2} \right) \]

\[ S = e^{-\frac{d'(X,Y)^2}{\sigma^2}} \]

\[ D = \sum_{j} S_{ij} \]

\[ \Delta = I - D^{-1/2} SD^{-1/2} \]
Major clusters correspond to the "outlier" eigenvalues.
#(Clusters)=20

Take a guess!
Sectors

1 Basic Materials
2 Conglomerates
3 Consumer Goods
4 Financial
5 Healthcare
6 Industrial Goods
7 Services
8 Technology
9 Utilities

Guess?
The Recent “Battle”
Notice the circle: A manifold, not a “ball” cluster.
The Cycle

[Diagram showing the cycle between different sectors: Financial, Services, Industrial Goods, Utilities, and Healthcare.]

Financial Services
Industrial Goods
Healthcare
Utilities
• Manifold Learning: Loads of current research!

• “Most Fun”=**persistent homology**. Detects the emergence of topology. (See: *Topology for Computing* by Zomorodian)

“Bottom up” method
Empty Sphere Method

Take a bunch of point...
Empty Sphere Method

Search for empty spheres...
Empty Sphere Method

Search for empty spheres...
Empty Sphere Method

Search for empty spheres...
Empty Sphere Method

Search for empty spheres...
Empty Sphere Method

Search for empty spheres...
Empty Sphere Method

Search for empty spheres...
Empty Sphere Method

Search for empty spheres...
Empty Sphere Method

Search for empty spheres...
Empty Sphere Method

Search for empty spheres...
Empty Sphere Method

Search for empty spheres...
Empty Sphere Method

Search for empty spheres...
Empty Sphere Method

Search for empty spheres...
Empty Sphere Method

Search for empty spheres...
Empty Sphere Method

Search for empty spheres...
Empty Sphere Method

Search for empty spheres...
Empty Sphere Method

Search for empty spheres...
Empty Sphere Method

Search for empty spheres...
Empty Sphere Method

Search for empty spheres...
Empty Sphere Method

Search for empty spheres...
Empty Sphere Method

Search for empty spheres...
The Delaunay Decomposition

Viola, le Delaunay Decomposition.
“Top Down”: Spectral Clustering

1. Form the Laplacian

2. List the K eigenvectors corresponding to the K smallest non-zero eigenvalues. (This is our dimension reduction into K-dimensional Euclidean space).

3. Normalize each row so all our points lie on a sphere.

4. Apply k-means.

Classical K-means Solution
Spectral Clustering

\[ L = \sum_{k=1}^{N} \sum_{x_i \in C_k} d(x_i, \mu_k) \approx 488 \]
Himalayas
Watch Out! 100 convergent runs of the k-means algorithm were performed.

\[ L = \sum_{k=1}^{N} \sum_{x_i \in C_k} d(x_i, \mu_k) \approx 536 \]

\[ L = \sum_{k=1}^{N} \sum_{x_i \in C_k} d(x_i, \mu_k) \approx 538 \]

\[ L = \sum_{k=1}^{N} \sum_{x_i \in C_k} d(x_i, \mu_k) \approx 534 \]
The Spectral Market

1 Basic Materials
2 Conglomerates
3 Consumer Goods
4 Financial
5 Healthcare
6 Industrial Goods
7 Services
8 Technology
9 Utilities

Guess who that is?
Another popular “bottom up” method

* - Linkage Clustering
  *=single-linkage, complete-linkage, average-linkage clustering....

Why are we looking all these different types of clustering procedures!....

No Free Lunch Theorem: Each reasonable cluster algorithm will be optimal/sub-optimal under the appropriate conditions.  (KNOW YOUR SYSTEM!)
Tree of Life

A dendrogram
1. Start by assigning each item to a cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distances (similarities) between the clusters the same as the distances (similarities) between the items they contain.

This discussion was taken from the very nice online tutorial: http://www.elet.polimi.it/upload/matteucc/Clustering/tutorial_html
2. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one cluster less.
3. Compute distances (similarities) between the new cluster and each of the old clusters.

We need to choose a measure of closest!
4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size N. (*)

<table>
<thead>
<tr>
<th></th>
<th>BA</th>
<th>FI</th>
<th>MI/TO</th>
<th>NA/RM</th>
</tr>
</thead>
<tbody>
<tr>
<td>BA</td>
<td>0</td>
<td>662</td>
<td>877</td>
<td>255</td>
</tr>
<tr>
<td>FI</td>
<td>662</td>
<td>0</td>
<td>295</td>
<td>268</td>
</tr>
<tr>
<td>MI/TO</td>
<td>877</td>
<td>295</td>
<td>0</td>
<td>564</td>
</tr>
<tr>
<td>NA/RM</td>
<td>255</td>
<td>268</td>
<td>564</td>
<td>0</td>
</tr>
</tbody>
</table>
4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size N. (*)

<table>
<thead>
<tr>
<th></th>
<th>BA/NA/RM</th>
<th>FI</th>
<th>MI/TO</th>
</tr>
</thead>
<tbody>
<tr>
<td>BA/NA/RM</td>
<td>0</td>
<td>268</td>
<td>564</td>
</tr>
<tr>
<td>FI</td>
<td>268</td>
<td>0</td>
<td>295</td>
</tr>
<tr>
<td>MI/TO</td>
<td>564</td>
<td>295</td>
<td>0</td>
</tr>
</tbody>
</table>
4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size N. (*)
Our Market...
The S&P500

S&P500 Dendogram
Sector break down of linkage clusters
If we apply it to our "manifold"

\[ d_{Average}(C_k, C_l) = \frac{\sum_{x \in C_k, y \in C_l} d(x, y)}{|C_k||C_l|} \]
Mat Lab

$Dist = \sin\left(\frac{\text{acos}(\text{Cor})}{2}\right)$;

$Z = \text{linkage}(\text{squareform}(\text{Dist}), \text{'average'});$;

$[H, T] = \text{dendrogram}(Z, 20);$;

$\text{Clusters} = \text{cluster}(Z, 20);$;

The choice of “closeness” is given by the average distance, called \textit{average-linkage clustering}.

\textit{Single-linkage clustering} uses the minimal distance....
Single-Linkage Clustering

$$d_{Single}(C_k, C_l) = \min_{x \in C_k, y \in C_l} d(x, y)$$
Complete-Linkage Clustering

\[ d_{\text{Complete}}(C_k, C_l) = \max_{x \in C_k, y \in C_l} d(x, y) \]
Combinations and Validation

- Validate clusters
- Use distinct clustering tools and explore if and how they disagree
- Combining Clustering: K-means for initial, then hierarchical methods to clean them up. This will work on very large data sets.
When we remove the clusters....

before

after
Fun Project!

- Explain the next scale’s “partition”.
- Devise ways to study discretely over lapping clusters. (Standard ways are Gaussian mixture models and fuzzy clustering type techniques).
Other Methods to study the network

• Fix a correlation level, retain edges with at least this much correlation, and study the resulting network.

• Small World: Clustering Coefficient and mean shortest path length.

• Degree distribution and power laws.
NYSE: (0.7:0.5:0.2)
NYSE