Detecting intraday financial market states using temporal clustering

Dieter Hendricks  Tim Gebbie  Diane Wilcox

QuERILab
Quantifying Emergence, Risk and Information
in financial markets

School of Computer Science and Applied Mathematics
University of the Witwatersrand, Johannesburg, South Africa

AASS Workshop, IMPA, 2016
Outline

1. Market microstructure and state representation

2. Identifying temporal system states
   - Unsupervised clustering approach
   - Giada-Marsili proposition with Noh ansatz
   - High-speed cluster configuration identification
   - Financial market intraday times as objects

3. Data and Results
   - Data
   - Temporal states, State Signature Vectors, Cluster size power-law fit, Estimated states, Transition probabilities

4. Conclusion
Market microstructure and state representation

- Financial markets as **complex adaptive systems** (Wilcox & Gebbie, 2015), prices and volumes as **measurable quantities**
- Market microstructure studies the **system evolution, behaviour** and consequences for **price formation** at the **lowest scale** (tick-level)
- Motivated by need for **efficient state representation** for automated trading agents in high-frequency markets to enable effective learning
- Agents faced with **high-volume, asynchronous** stream of financial market **data** from a real-time datafeed
- Can we find **persistent structure** in this streaming data, such that **meaningful learning** can take place for adaptive trading algorithms?
1 Market microstructure and state representation

2 Identifying temporal system states
   - Unsupervised clustering approach
   - Giada-Marsili proposition with Noh ansatz
   - High-speed cluster configuration identification
   - Financial market intraday times as objects

3 Data and Results
   - Data
   - Temporal states, State Signature Vectors, Cluster size power-law fit, Estimated states, Transition probabilities

4 Conclusion
Identifying temporal system states
Unsupervised clustering approach

- Clustering involves grouping objects according to metadata describing objects or their associations.
- For unsupervised clustering, apply super-paramagnetic ordering of q-state Potts model for cluster identification (Blatt et al., 1996).
- Cost function is a Hamiltonian whose low energy state corresponds to a cluster configuration most compatible with the sample.

\[ H_g = - \sum_{s_i, s_j \in S} J_{ij} \delta(s_i, s_j) - \frac{1}{\beta} \sum_i h_i^M s_i \]

where spins \( s_i \) can take on \( q \)-states and the short-range ferromagnetic interaction between spins is given by \( J_{ij} \).

- To parameterise the model, make Noh (2000) ansatz and use this to develop the MLE approach of Giada & Marsili (2001).
Outline

1. Market microstructure and state representation

2. Identifying temporal system states
   - Unsupervised clustering approach
   - Giada-Marsili proposition with Noh ansatz
   - High-speed cluster configuration identification
   - Financial market intraday times as objects

3. Data and Results
   - Data
   - Temporal states, State Signature Vectors, Cluster size power-law fit, Estimated states, Transition probabilities

4. Conclusion
Objects belonging to the same cluster share a common component,
\[ \tilde{x}_i = g_{si} \bar{\eta}_{si} + \sqrt{1 - g_{si}^2} \bar{\epsilon}_i. \] (1)

If we take this as a statistical hypothesis, and assume \( \bar{\eta}_{si} \) and \( \bar{\epsilon}_i \) are Gaussian, this leads to the following probability density,
\[
P(\{\tilde{x}_i\}|\mathcal{G}, S) = \prod_{d=1}^{D} \left( \prod_{i=1}^{N} \delta \left( x_i(t) - g_{si} \bar{\eta}_{si} + \sqrt{1 - g_{si}^2} \bar{\epsilon}_i \right) \right) \] (2)

and the maximum likelihood for structure \( S \) can be written as (Giada & Marsili, 2001),
\[
\mathcal{L}_c(S) = \frac{1}{2} \sum_{s:n_s>1} \left( \log \frac{n_s}{c_s} + (n_s - 1) \log \frac{n_s^2 - n_s}{n_s^2 - c_s} \right). \] (3)

Full derivation in Hendricks, Gebbie & Wilcox (2016b)
1 Market microstructure and state representation

2 Identifying temporal system states
   - Unsupervised clustering approach
   - Giada-Marsili proposition with Noh ansatz
   - High-speed cluster configuration identification
   - Financial market intraday times as objects

3 Data and Results
   - Data
   - Temporal states, State Signature Vectors, Cluster size power-law fit, Estimated states, Transition probabilities

4 Conclusion
Identifying temporal system states
High-speed cluster configuration identification

- Likelihood function in Equation 3 used as **objective function** in metaheuristic optimisation routine (genetic algorithm)
- Systematically **evaluate** candidate **configurations** \( (S) \), converging towards **best approximation** of data descriptor
- Hendricks, Gebbie & Wilcox (2016a) demonstrate a high-speed **Parallel Genetic Algorithm** implementation in **CUDA**, using the SPMD architecture to enumerate the GPU thread hierarchy with population members for **concurrent application of genetic operators**
- A **low-cost**, **scalable**, **high-speed** implementation for **unsupervised cluster detection**
Outline

1. Market microstructure and state representation

2. Identifying temporal system states
   - Unsupervised clustering approach
   - Giada-Marsili proposition with Noh ansatz
   - High-speed cluster configuration identification
   - Financial market intraday times as objects

3. Data and Results
   - Data
   - Temporal states, State Signature Vectors, Cluster size power-law fit, Estimated states, Transition probabilities

4. Conclusion
Generic data generative model ansatz conducive to any problem where Gaussian object and cluster innovations are reasonable.

Marsili (2002) grouped days according to closing price performance to identify temporal states.

Propose that intraday temporal regimes of financial markets are characterised by feature performance of stocks.

Using trade price, trade volume, spread and volume imbalance features of TOP40 stocks on the JSE, we find clusters of 60-min, 30-min, 15-min and 5-min periods.

Determine whether clustering at varying calendar time scales reveals interesting hierarchy of system behaviour.

Reduces significant amount of high-frequency information into a tractable representation for intraday learning.
Identifying temporal system states

State Signature Vectors

- Temporal clustering reveals *ex-ante* grouping of periods
- How do we incorporate this into an **online learning algorithm**? i.e. determine the state we are **currently** in
- Analyse **average feature performance** of stocks within identified temporal clusters and extract *State Signature Vector (SSV)* as state descriptor
- Online feature vectors are **easy to compute** from streaming market datafeeds in financial markets, conducive to near-real-time detection
- Need to determine which identified states are **significant**, i.e. likely to persist such that meaningful learning is possible
- Use best **power law fit** to cluster size to identify likely persistent states
- Use **Euclidean distance** of (current) online feature vector to set of SSVs for state index assignment

Dieter Hendricks  Tim Gebbie  Diane Wilcox (Detecting intraday financial market states usi) AASS Workshop, IMPA, 2016
Outline

1. Market microstructure and state representation

2. Identifying temporal system states
   - Unsupervised clustering approach
   - Giada-Marsili proposition with Noh ansatz
   - High-speed cluster configuration identification
   - Financial market intraday times as objects

3. Data and Results
   - Data
     - Temporal states, State Signature Vectors, Cluster size power-law fit, Estimated states, Transition probabilities

4. Conclusion
Data and Results

Data

- Tick-level trades and top-of-book quotes for 42 stocks on the Johannesburg Stock Exchange (JSE) from 1 November 2012 to 30 November 2012
- Raw data sourced from Thomson Reuters Tick History
- Stored in MongoDB noSQL database on QuERILab server and integrated into MATLAB development environment
Outline

1 Market microstructure and state representation

2 Identifying temporal system states
   - Unsupervised clustering approach
   - Giada-Marsili proposition with Noh ansatz
   - High-speed cluster configuration identification
   - Financial market intraday times as objects

3 Data and Results
   - Data
     - Temporal states, State Signature Vectors, Cluster size power-law fit, Estimated states, Transition probabilities

4 Conclusion
Identifying temporal system states
60-min clusters (TOP40 stocks; price, volume, spread, volume imbalance),
01-Nov-2012 to 30-Nov-2012, GREEN=morning, YELLOW=lunch, RED=afternoon

Figure: JSE TOP40 60-minute temporal clusters for the period 01-Nov-2012 to 30-Nov-2012, representing 184 distinct periods. Each node represents a 60-minute period during a trading day, with the colour shading indicating the time-of-day (Morning = green, Lunch = yellow, Afternoon = red) and node connectedness indicating cluster membership.
Test for **cluster size** power law fit: *60-min periods*

Power law distributional form \((p(x) \sim x^{-\alpha})\) vs empirical data

\(\alpha = 6.35\), \(p_{value} = 0.97637\), \(x_{min} = 13\), \(L = -12.2752\)
State Signature Vectors

60-min clusters (TOP40 stocks; price, volume, spread, volume imbalance), 01-Nov-2012 to 30-Nov-2012

Cluster 5 (21 members, $c_5 = 35483.445$)

Cluster 6 (15 members, $c_6 = 7323.8035$)

Cluster 7 (16 members, $c_7 = 542.8251$)

Cluster 12 (14 members, $c_{12} = 982.9791$)

Cluster 13 (13 members, $c_{13} = 331.5971$)

Cluster 15 (13 members, $c_{15} = 579.9487$)
Estimating temporal states using feature vectors
60-min clusters (TOP40 stocks; price, volume, spread, volume imbalance), 01-Nov-2012 to 30-Nov-2012
Actual vs Estimated temporal states
60-min clusters (TOP40 stocks; price, volume, spread, volume imbalance),
01-Nov-2012 to 30-Nov-2012

(a) Actual 60-min states
(b) Estimated 60-min states
Identifying temporal system states
30-min clusters (TOP40 stocks; price, volume, spread, volume imbalance),
01-Nov-2012 to 30-Nov-2012, GREEN=morning, YELLOW=lunch, RED=afternoon
Test for **cluster size** power law fit: *30-min periods*

Power law distributional form ($p(x) \sim x^{-\alpha}$) vs empirical data

$\alpha = 9.91$, $p_{value} = 0.86167$, $x_{min} = 14$, $L = -15.5522$
State Signature Vectors
30-min clusters (TOP40 stocks; price, volume, spread, volume imbalance), 01-Nov-2012 to 30-Nov-2012

Cluster 4 (15 members, $c_4 = 260.3446$)

Cluster 7 (15 members, $c_7 = 593.0322$)

Cluster 9 (15 members, $c_9 = 668.719$)

Cluster 10 (14 members, $c_{10} = 378.3076$)

Cluster 11 (21 members, $c_{11} = 34233.7002$)

Cluster 13 (14 members, $c_{13} = 363.4887$)

Cluster 18 (14 members, $c_{18} = 307.6546$)

Cluster 19 (14 members, $c_{19} = 555.1054$)

Cluster 25 (14 members, $c_{25} = 388.596$)

Cluster 28 (17 members, $c_{28} = 767.8294$)
Estimating temporal states using feature vectors
30-min clusters (TOP40 stocks; price, volume, spread, volume imbalance),
01-Nov-2012 to 30-Nov-2012
Actual vs Estimated temporal states
30-min clusters (TOP40 stocks; price, volume, spread, volume imbalance), 01-Nov-2012 to 30-Nov-2012

(a) Actual 30-min states  
(b) Estimated 30-min states
Identifying temporal system states

5-min clusters (TOP40 stocks; price, volume, spread, volume imbalance), 01-Nov-2012 to 30-Nov-2012, GREEN=morning, YELLOW=lunch, RED=afternoon
Stability of the online state assignment out-of-sample \textit{ex-ante} (01-Nov-2012 to 30-Nov-2012, same period used for SSV estimation) vs \textit{ex-post} (03-Dec-2012 to 07-Dec-2012, one week after SSV estimation window)

Boxplot of Euclidean distance of best-match 30-min state assignments \textit{ex-ante} vs \textit{ex-post}

Boxplot of Euclidean distance of best-match 5-min state assignments \textit{ex-ante} vs \textit{ex-post}

(a) 30-min states

(b) 5-min states
Conclusion

- Provided a scheme for **unsupervised** determination of intraday financial market **states** and **online** state **detection** using **SSVs**
- Tractable **state representation** for learning agents in high-frequency trading domain
- Non-trivial **hierarchy** of clustering **behaviour** at varying time-scales reveals **scale-specific information** for learning
- One application is the online enumeration and refinement of an **empirical transition probability matrix**
- States appear to persist **long enough** to be exploited - needs to be verified
- Will consider clustering in **machine time** in further work
For Further Reading I


