Market state discovery

by Unarine Singo

June 2020

Content

- What is market state discovery?
- Project objectives
- Introduction to Inverse Covariance Clustering (ICC)
- Introduction to Agglomerative Super-Paramagnetic Clustering (ASPC)
- Reverse ASPC (Noh Anzats) data simulation
- Next steps and questions
- References

What is market state discovery?

- Market state discovery is synonymous to unsupervised multivariate clustering; however, we endeavour to cluster market observations in time
- These different 'states' of the market are commonly attributed in literature to unobservable, or latent, regimes representing a set of macroeconomic, market sentiment value
- State discovery has been studied under different names. Some examples include regime-switching (Hamilton, 2005), temporal clustering (Hendricks, Gebbie, Wilcox, 2016) and volatility modelling (Bollerslev, Chou, Kroner, 1992). In general, all methods aim to find a lower-dimensional representation of the complex structure of a financial market
- Bull and Bear terminology are simple examples of how market participants have developed their own language to summarise market information into discrete 'states'
- Whether market states exist and are persistent in time is an open question. However, some interesting papers have proposed a structured methodology for identifying them
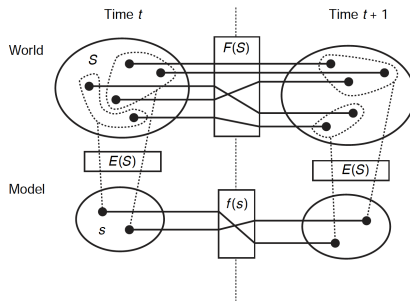
Figure 3.2. A formal model of models. Dotted lines indicate the equivalence class mappings.

Figure: Excerpt from *Adaptive Complex Systems* by Scott E. Page

"...To reduce the size of the state space, the modeler generates equivalence classes: maps from a subset of the real-world states, S, to a model state, s. Here we designate the equivalence class map as $E(S)$. Based on these new model states, the quest of the scientist is to find a useful transition function, $f(s)$, for the model."

- Compare Inverse Covariance Clustering (ICC) (Aste, Procacci, 2019) to Agglomerative Super-Paragmetic Clustering (ASPC) (Gebbie, Yelibi, 2019)
  - Apply both methods to low dimensional synthetic datasets (using reverse ASPC) to explore their robustness
  - Scale up to higher dimensional datasets of daily equity returns of the JSE and explore detected states
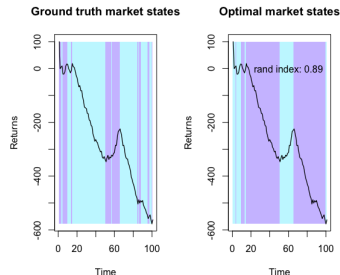- Test persistence of market states though supervised learning on the states detected by ICC

Figure: Preliminary performance results of ICC. The algorithm is implemented on a simulated joint Gaussian 2-stock market

Introduction to Inverse Covariance Clustering

Aste and Procacci (2019) proposed Inverse Covariance Clustering, an unsupervised learning approach to identifying clusters of financial time-series. It's novelty is two fold:

- It can with computational efficiency identify states of high dimensional multivariate data-sets
- The algorithm has an additional hyper-parameter, termed the temporal-cohesion parameter, which penalises frequent state switching
- ICC is grounded on the assumption that market state can be discovered by a precision matrix and vector of expectation values which are associated with a set of multivariate observations clustered together accordingly with a given procedure. The method is comprised of two steps, **segmentation** (initialisation) and **cluster assignment** (optimisation)

Segmentation

- Begins by setting the number of clusters K and assigns multivariate observations to clusters randomly
- Each multivariate observation, $X_t = [x_{t,1}, x_{t,2}, x_{t,3}, ..., x_{t,N}]$, is associated with a single day's return of N stocks
- From these K sets of data we compute sample means $\vec{\mu_k}$ and precision matrices $\mathbf{J_k}$ and then iteratively re-assign points to the cluster with the smallest

$$M_{t,k} = d_{t,k}^2 + \gamma 1\{K_{t-1} \neq 1\} \tag{1}$$

- where $d_{t,k}^2 = (X_t - \vec{\mu_t})^T J_k(X_t - \vec{\mu_t})$ is the square Mahalanobis distance of observation X_t in state k with respect to the state centriod $\vec{\mu_k}$. γ is the temporal cohesion parameter that penalises frequent state switching

Cluster assignment

- State re-assignment is made computationally by efficient by using the Viterbi algorithm and sparsification of the precision matrix.
- The Viterbi algorithm transforms the distance computation from polynomial-time $O(K^T)$ to linear time complexity $O(KT)$
- Sparsification of the precision matrix, J_k, significantly reduces the number of parameter computations from N^2, and the precision matrix is computationally made efficient by using a network filtering approach called TMFG-LoGo

The figure below provides a visualisation of the problem of assigning points to clusters. Based on the parameter estimates, we compute the Mahalanobis distance of every multivariate observation obtaining and each cluster k. The Viterbi algorithm solves for the state / cluster path with the minimum total distance.
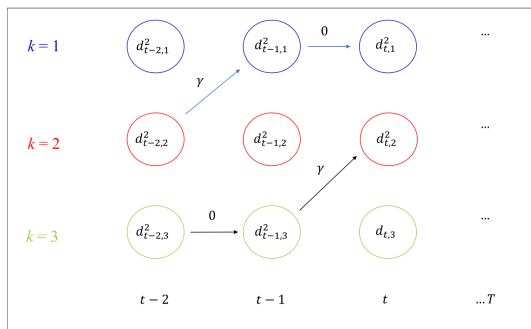
Figure: Example of two among the K^T possible paths considering $K = 3$ clusters and T observations. If an observation is assigned to same cluster as the previous one, no penalty is applied, otherwise a cost weighted by the parameter γ is added.

Introduction to Inverse Covariance Clustering

Preliminary replication results on varying Markov Chain dependent state sequences
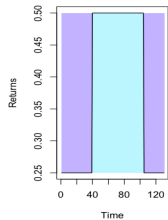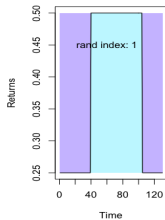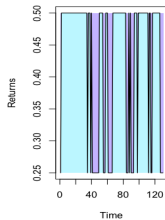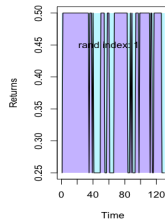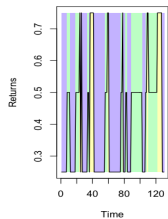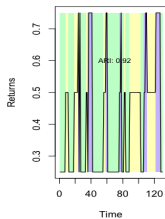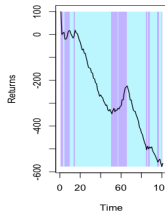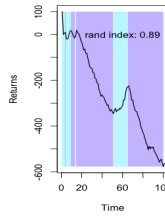
- Gebbie and Yelibi (2019) proposed ASPC, a fast and universal clustering method that is analogous to a Potts Model
- The method is a computational improvement based on Giada-Marsili (2001) likelihood model
- ASPC can be used to cluster stocks (horizontal clustering) and states (vertical clustering)
- ASPC can also be used to generate synthetic time-series datasets (Reverse ASPC)

Agglomerative Fast Super-Paramagnetic Clustering

Lionel Yelibi[*] and Tim Gebbie[†]

Department of Statistical Science, University of Cape Town, Rondebosch, South Africa
(Dated: August 5, 2019)

We consider the problem of fast time-series data clustering. Building on previous work modeling the correlation-based Hamiltonian of spin variables we present a fast non-expensive agglomerative algorithm. The method is tested on synthetic correlated time-series and noisy synthetic data-sets with built-in cluster structure to demonstrate that the algorithm produces meaningful non-trivial results. We argue that ASPC can reduce compute time costs and resource usage cost for large scale clustering while being serialized and hence has no obvious parallelization requirement. The algorithm can be an effective choice for state-detection for online learning in a fast non-linear data environment because the algorithm requires no prior information about the number of clusters.

Figure: Abstract from the ASPC paper on arxiv.org

- Consider a financial market data set $\Xi = \{\vec{\xi_i}\}_{i=1}^{N}$ composed of N sets of $\vec{\xi_i} = \{\xi_i(\vec{d})\}_{d=1}^{D}$
- For example, when clustering stocks, $\xi_i(d)$ is the normalised daily returns of asset i in day d
- For simplicity, we assume $\xi_i(d)$ are Gaussian variables
- We invoke a statistical anzats proposed by Noh that assumes stocks can be decomposed into a cluster related effect and a random effect:

$$\xi_i(d) = g_{s_i}\eta_{s_i}(d) + \sqrt{1 - g_{s_i}^2}\epsilon_i(d) \tag{2}$$

- Where g_{s_i} is the intra-cluster coupling parameter, s_i, are integer cluster variables (so-called Potts Spins), η_{s_i}, the cluster related influence and ϵ_i the random effect
- In order to fit the dataset Ξ with equation 2, we introduce the Dirac-delta function and compute the likelihood of the data with structure $\mathcal{S} = \{s_i\}_{i=1}^{N}$ and parameters $\mathcal{G} = \{g_s\}_{s=1}^{N}$

$$P = \prod_{d=1}^{D}\prod_{i=1}^{N}\left\langle \delta\left(\xi_i(d) - \left(g_{s_i}\eta_{s_i}(d) + \sqrt{1 - g_{s_i}^2}\epsilon_i(d)\right)\right)\right\rangle \tag{3}$$

- The joint liklihood is the probability of a cluster configuration matching observed data
- The resultant log-liklihood from equation 3 can thought of as a Potts Hamiltonian:

$$L_c = \frac{1}{2} \sum_{s:n_s>1} \ln \frac{n_s}{c_s} + (n_s - 1) \ln \frac{n_s^2 - n_s}{n_s^2 - c_s} \tag{4}$$

- ASPC proposes a agglomerative bottom-up implementation of Giada-Marsili's merging algorithm to optimise equation 4

- The Noh Anzats in equation 2 can also be used to simulate correlated time-series using the following equivalent equation:

$$\xi_i(d) = \frac{\sqrt{g_{s_i}}\eta_{s_i}(d) + \epsilon_i(d)}{\sqrt{1 + g_{s_i}}} \tag{5}$$

- When identifying or simulating states, objects, $i = \{1, 2, 3, ..., N\}$ are days and features $d = \{1, 2, 3, ..., D\}$ are stocks
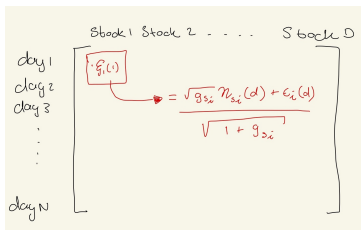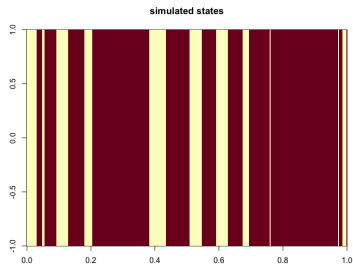
Figure: Step 1: Define values for number of clusters, C, size of clusters, s and obtain $N = s * C$ the number of days in the dataset. Pick the number of stocks D. Then based on a Markov Chain, create sequence of spin-labels
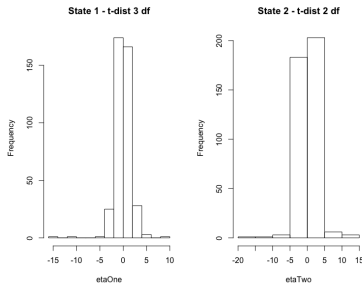
Figure: Step 2: Create a CxD array of state effects η_{s_i}
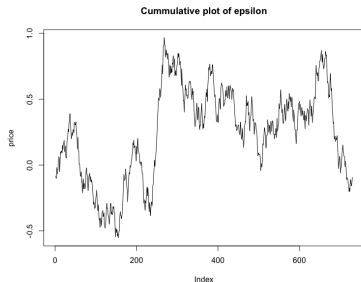
Figure: Step 3: Create a NxD array of random effects $\epsilon \sim \mathcal{N}(0, 1)$

Figure: Step 4: Create NxD array and compute daily returns using the Noh anzats in equation 5

and fix intra-cluster binding strength g_s (which is set to 0.1)
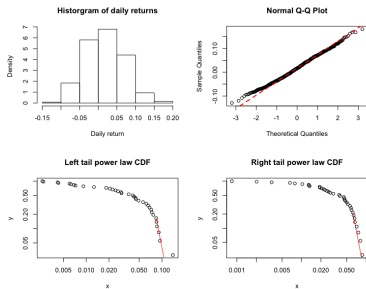
Figure: Stylised facts of simulated data set
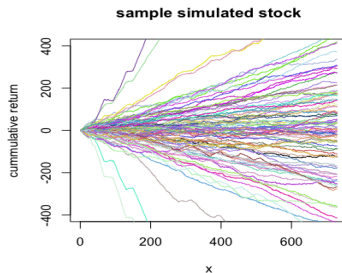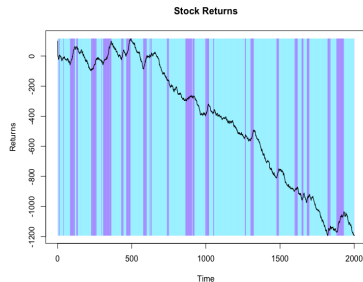

Figure: Sample of 100 simulated stocks
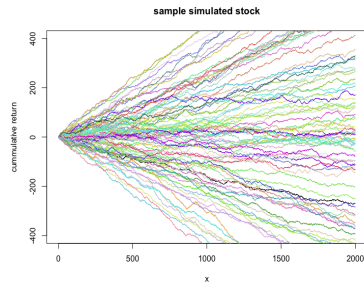
Figure: Simulated daily market return

Figure: Sample of 100 simulated stocks

- Finalise reverse ASPC data simulation
  - What are sensible distributions for η_{s_i}?
  - How many data sets are sufficient for comparison? (e.g., varying states numbers, number of stock, number of days)
- Apply ICC and ASPC to simulated data set and measure performance using Adjusted Rand Index
- Apply ICC and ASPC to real data set and explore identified states
- Test market state persistence by means of supervised learning on identified states

Authur, B., (1995), Complexity in Economic and Financial Markets, Complexity, 1, 20-25, 1995

Hamilton, J. D. (2005). Regime Switching Models., https://doi.org/10.1201/9781315373751-9

D. Hendricks, T. Gebbie and D. Wilcox (2016) Detecting intraday financial market states using temporal clustering, Quantitative Finance, 16:11, 1657-1678, DOI: 10.1080/14697688.2016.1171378

Bollerslev, T., Chou, R. Y., and Kroner, K. F. (1992)., ARCH modeling in finance. A review of the theory and empirical evidence., Journal of Econometrics, 52(1–2), 5–59. https://doi.org/10.1016/0304-4076(92)90064-X

Wilcox, D., and Gebbie, T. (2015)., Hierarchical Causality in Financial Economics., SSRN Electronic Journal, 1–16. https://doi.org/10.2139/ssrn.2544327

Miller, J. H., and Page, S. E. (2009)., Complex adaptive systems: An introduction to computational models of social life. In Complex Adaptive Systems: An Introduction to Computational Models of Social Life., https://doi.org/10.1080/01488370802162558

Blatt, M., Wiseman, S., and Domany, E. (1996)., Superparamagnetic clustering of data., Physical Review Letters, https://doi.org/10.1103/PhysRevLett.76.3251

Giada, L., and Marsili, M. (2001). Data clustering and noise undressing of correlation matrices. Physical Review E - Statistical, Nonlinear, and Soft Matter Physics, 63(6), 1–8. https://doi.org/10.1103/PhysRevE.63.061101

Yelibi, L., Gebbie, T. (2019). Agglomerative Fast Super-Paramagnetic Clustering. (2). Retrieved from http://arxiv.org/abs/1908.00951

Procacci, P. F., Aste, T. (2019). Forecasting market states. Quantitative Finance, 19(9), 1491–1498. https://doi.org/10.1080/14697688.2019.1622313

The End