Simulating Price Interactions by Mining Multivariate Financial Time Series

Bruno Silva
DSI/ESTSetúbal
Instituto Politécnico de Setúbal
Portugal
bruno.silva@estsetubal.ips.pt

Luis Cavique
Universidade Aberta
Portugal
lcavique@univ-ab.pt

Nuno Marques
DI/FCT - UNL
Portugal
nmm@di.fct.unl.pt

Abstract

This position paper proposes a framework based on a feature clustering method using Emergent Self-Organizing Maps over streaming data (UbiSOM) and Ramex-Forum – a sequence pattern mining model for financial time series modeling based on observed instantaneous and long term relations over market data. The proposed framework aims at producing realistic *monte-carlo* based simulations of an entire portfolio behavior over distinct market scenarios, obtained from models generated by these two approaches.

1 Introduction

Grasping the apparently random nature of financial time series has proven to be a difficult task and countless methods of forecasting are presented in literature. Nowadays, this is even more difficult due to a global economy with strong interconnections. Most traders forecast future price using some combination of fundamentals, indicators, patterns and experience in the expectation that recent history will forecast the probable future often enough to make a profit. Detecting correlations between financial time series and being able to simulate both short and long term interactions in virtual scenarios using models extracted from observed market data can provide an increasingly needed tool to minimize risk exposure and volatility for a given portfolio of securities. This position paper argues that feature clustering methods using Emergent Self-Organizing Maps over streaming data (UbiSOM) [Silva *et al.*, 2012], [Silva and Marques, 2010b] can be conjoined with Ramex-Forum – a sequence pattern mining model [Marques and Cavique, 2013], for financial time series modeling based on observed instantaneous and long term relations over market data. Since the lower the correlation among the individual securities, the lower the overall volatility of the entire portfolio, this makes possible to propose a tool to minimize risk exposure and volatility for a given portfolio of securities. The proposed framework aims at producing more realistic *Monte Carlo*-based simulations of the entire portfolio behavior over distinct market scenarios, obtained from models generated by these two approaches.

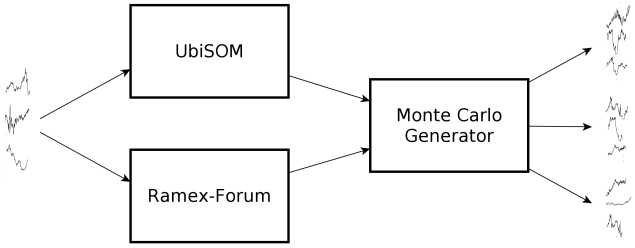

Figure 1: Proposed framework (description in text).

2 Proposed Framework

The proposed modular framework is depicted in Figure 1 and consists of i) The UbiSOM, an ESOM algorithm tailored from streaming data ii) The Ramex-Forum, a sequence pattern mining model and iii) A *Monte Carlo*-based simulator. The first two are fed with a stream of log-normalized raw asset prices, which are then used by the third module to produce future different and possible market scenarios, based on the observed data. The UbiSOM can model instantaneous short-term correlations between the various assets and its topological map (Section 2.1) can be used as a starting point to generate alternate time-series based on a trajectory model (Section 3.1) by the simulation module. The input from the Ramex-Forum module should be useful to incorporate in the simulations long-term dependencies between the assets to produce more realistic market scenarios.

2.1 Emergent Self-Organizing Maps

Self-Organizing Maps [Kohonen, 1982] can use the ability of neural networks to discover nonlinear relationships in input data and to derive meaning from complicated or imprecise data for modeling dynamic systems such as the stock market. The Self-Organizing Map (SOM) is a single layer feedforward network where the output neurons are arranged in a 2-dimensional lattice, whose neurons become specifically tuned to various input vectors (observations) in an orderly fashion. Each input x^d is connected to all output neurons and attached to each neuron there is a weight vector w^d with the same dimensionality as the input vectors. These weights represent *prototypes* of the input data. The topology preservation of the SOM projection is extensively used by focusing SOM on using larger maps – ESOM [Ultsch and Her-

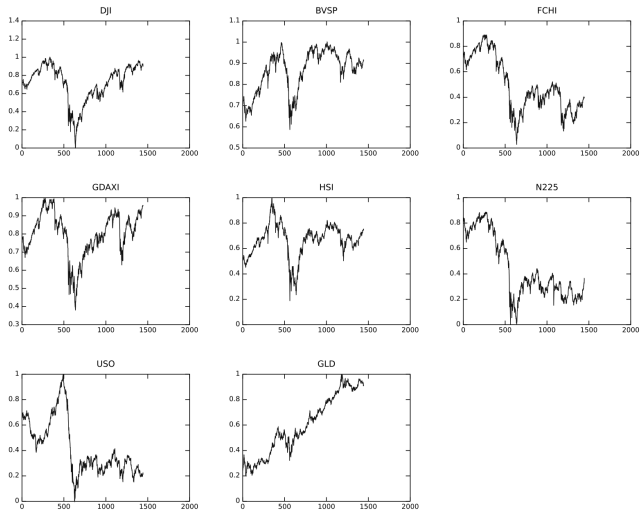

Figure 2: Time series (closing prices) for the described eight financial products from 2006 to 2012 (description in text). The values have been normalized to logarithmic scale.

rmann, 2005]. A previous work [Silva and Marques, 2010a] showed that ESOMs provide a way of representing multivariate financial data on two dimensions and are a viable tool to detect instantaneous short-term correlations between time-series. We illustrate this in Section 3, within our preliminary results. Additionally, and supported by the detected correlations, the topological ordered map can be used as a good starting point to generate realistic multivariate financial data based on the short-term relationships.

2.2 The Ramex-Forum Algorithm

Ramex-Forum solves the problem of huge number of rules that avoid a global visualization in many pattern discovery techniques (e.g., [Agrawal and Srikant, 1995]). Ramex-Forum is a sequential pattern mining algorithm that includes a two-phase; the transformation phase and the search phase. In the transformation phase the dataset is converted into a graph where cycles are allowed. The raw data must be sorted in such a way that each time interval can be identified. In the search phase the maximum weighted spanning poly-tree is found. A poly-tree is a direct acyclic graph with one path between any pair of nodes at the most. The in-degree of any vertex of a tree is 0 (the root) or 1. On the other hand, the in-degree of a vertex can be greater than 1. A maximum weighted spanning poly-tree is the spanning poly-tree with a weight that is upper than or equal to the weight of every other spanning poly-tree. The Ramex-Forum algorithm develops a new heuristic inspired in Prim's algorithm [Prim, 1957] and assures a new way of visualization long term patterns in polynomial execution time.

3 Preliminary Results

In this section we provide a proof-of-concept of the proposed methodology within the framework. The proposed method is illustrated with historical data representing the world econ-

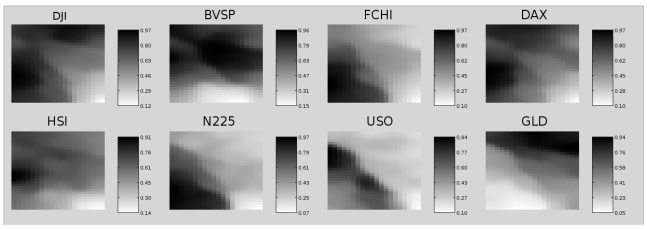

Figure 3: Component planes for studied time series (20×30 trained SOM). Similarities indicate correlated time series.

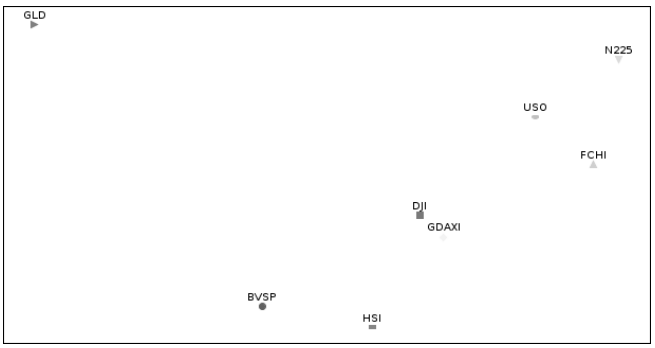

Figure 4: Projection of obtained distances between the component planes in Figure 3.

omy in the recent past (years 2006 to 2012) – Figure 2. The huge economic changes during this period are good to show the usefulness of data mining algorithms over financial data. Top financial products such as average Indexes for companies based in different countries (DJI – Dow Jones, in the United States; BVSP – Bovespa, in Brazil; FCHI, Euronext in Paris; N225, Nikkei in Japan; the HSI — Hang Seng Index, in Hong Kong; and DAX, German Index) and relevant commodities exchange-traded funds (ETF) such as United States Oil Fund (USO) and GLD for a physically backed gold ETF, were considered.

Each time series is considered a feature of the training data, i.e., observations are the prices of the financial products for consecutive days. After performing a logarithmic normalization of the values, so that the specific range of each asset price is disregarded, the historical data forms the training dataset that is fed into the UbiSOM and Ramex-Forum modules.

The trained UbiSOM map contains a topologically organized projection of the historical data. Correlations between individual time-series are extracted through a visualization technique for the UbiSOM map. By component plane representation we can visualize the relative component distributions of the input data. Component plane representation can be thought as a sliced version of the UbiSOM. Each component plane has the relative distribution of one data vector component. In this representation, dark values represent relatively large values while white values represent relatively small values. By comparing component planes we can see if two components correlate. If the outlook is similar, the components strongly correlate. The component planes for the resulting trained map are represented in Figure 3. Visual inspection

may suffice to detect correlations, but in [Silva and Marques, 2010a] we provided an automated algorithm to cluster time-series based on a distance metric computed for pairs of component planes (Figure 4). However, this ability is only of relevance in this paper to justify the use of ESOM maps to generate multivariate time-series based on a trajectory model (Section 3.1).

Figure 5 presents a Ramex-Forum generated graph for this financial data, considering interactions with a latency of up to 160 (long-term) trading days over a period of 2000 days – results presented in [Marques and Cavique, 2013]. Each arc represents the number of synchronous positive price tendencies (buying signals given by a moving average indicator). During the studied period Hong Kong HSI Index has a behavior that was preceded by 273 times by similar variations in American Dow Jones (DJI) and 179 times by German DAX. We should notice that USA DJI is influencing most major assets in the world. The only exception to this is European German DAX, that strongly co-influences Chinese HSI. Another correlation found is between HSI and GLD tendencies in these long term dependencies. This is something that UbiSOM cannot capture and can be incorporated when generating more realistic market scenarios.

3.1 Generating Scenarios

By projecting again the historical data over the UbiSOM we get a set of trajectories over the map that are used to generate these alternate time series. A trajectory is formed by projecting two consecutive observations from the training data and storing the pair of neurons that were activated, in the form of a trajectory (bare in mind that loops are frequent, because two similar observations are prone to be projected in the same neuron). Figure 6 depicts these trajectories in the form of a directed graph in which each vertex represents a neuron and the edges the obtained trajectories. The weight of the edge indicates how many times the trajectory was followed in the projection of the training data.

Based on this *trajectory model*, we can generate alternate time series using *Monte Carlo* simulations. Starting at a vertex with edges and randomly choosing the next trajectory of the model to follow we can create paths of arbitrary lengths (dependent of the desired number of daily prices). Each vertex/neuron contains the prototype of data that contains the set

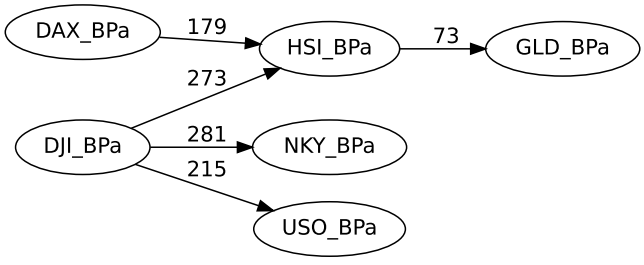

Figure 5: Correlation Ramex-Forum graph considering interactions with a latency of up to 160 trading days over a period of 2000 days.

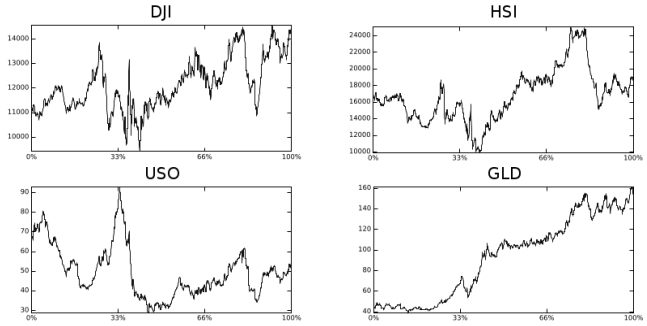

Figure 7: Simulated outcome of a possible market scenario from the trajectory model of a 15×10 trained UbiSOM.

of daily prices for similar observed days. The totality of the path then gives us the multivariate time series. Details on how to generate the path are currently being studied, it must not be totally random, i.e., the weight on the edges must be taken into account so as to give more importance to trajectories that are more common. Also, when creating the trajectory model we store at each vertex/neuron the statistical variation of the training vectors that are projected on that particular neuron. This allows generating a Gaussian around each prototype vector component to introduce variability on a particular virtual daily price. This is particularly important when loops are being followed in the path, so that generated time series doesn't contain "flat" lines.

Figure 7 depicts a sample of a generated outcome that can be obtained from trajectories over the trained UbiSOM. It can be seen that the multivariate time series maintain the observed correlations in the original historical data. This can be very useful in generating possible scenarios for risk estimation.

3.2 Discussion

Visual inspection of the similarity of component planes in Figure 3 shows that the results of the UbiSOM model are coherent, e.g., DJI and DAX are strongly correlated in their historical behavior. GLD, as it was expected, is very far from any other financial product. Its historical behavior is extremely different in the analysis period, mainly always in a upward movement. All the other assets maintain a significant distance from the others, showing that the correlation is not that strong. Interesting additional long term relations were found by Ramex-Forum algorithm. The example shown in Figure 5, presents the USA DJI index influencing most other indexes. Also China (HSI) is detected as a major player in world economy and seems to be the major influence on the price of gold (GLD). Indeed, during the analyzed period, People's Republic of China was one of the major buyers of gold in the world and has the largest reserves of Gold and Foreign Exchange in the world (CIA World Factbook, 2013).

However, it is the conjunction of the long term relations detected in Ramex-Forum with the magnitude of short time multivariate dependencies of UbiSOM, that should be the most interesting application. Different long term trajectories can now be generated on the UbiSOM map, based on the long term sequences detected by Ramex-Forum. E.g., neurons cor-

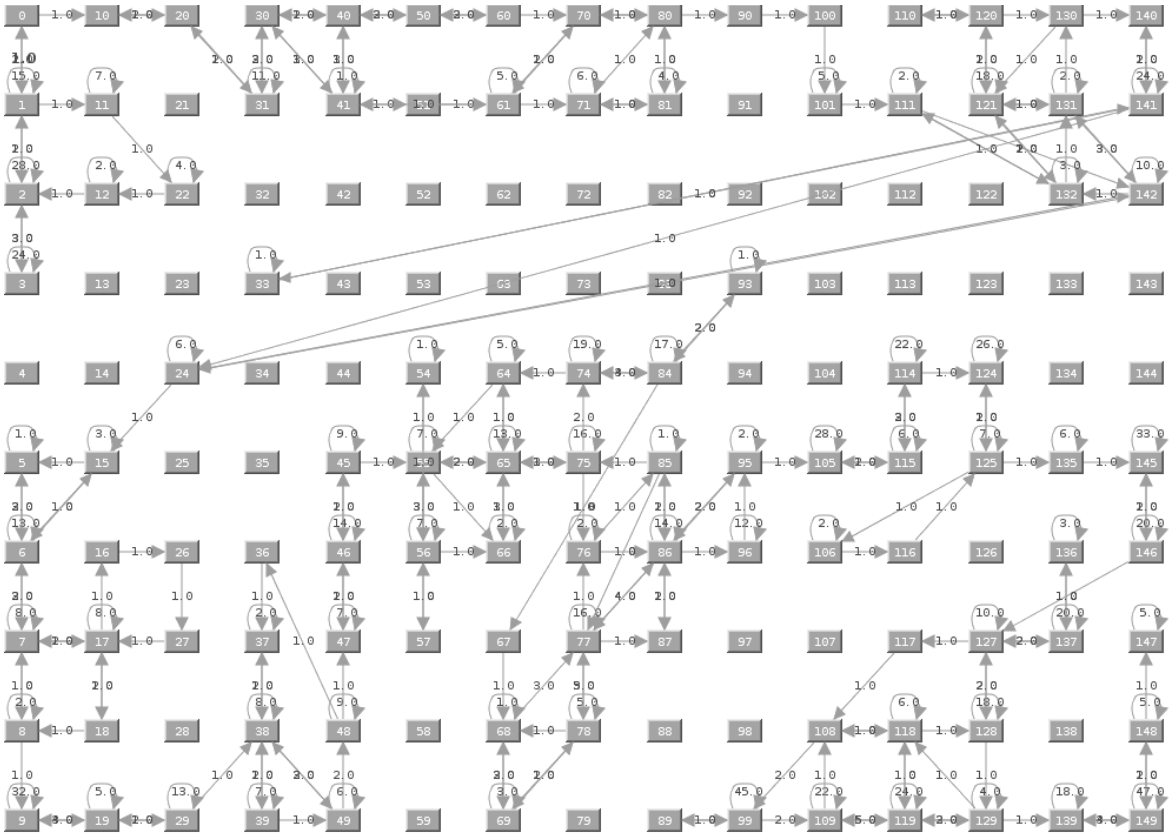

Figure 6: Trajectories generated for the projection of the historical training data over a 15×10 trained UbiSOM.

responding to highest increases in gold values can be easily selected from a SOM map. The same could be done for high values for the Chinese (HSI), or USA (DJI) economy. Highly probable pathways should then be made among those neurons. In practice these will encode the Ramex-Forum graph as a probable pathway among distinct UbiSOM neurons. Then for a given trading day (e.g. today) starting point, we can then generate random walks in the map. Since each neuron represents a possible market state, we can easily generate for each neuron a possible *virtual trading day* that is strongly related with observed data. However it will be the pathways to provide the most interesting effect on this map. Indeed, in average, virtual trading days will follow possible sequences given (and measured) by Ramex-Forum graph.

4 Conclusions

Both algorithms should provide a realistic and easily usable framework to study and simulate possible effects of either economic or political decisions. Even extreme events, e.g., *Acts of God*, may them be given some probability. We believe that such a time-series based model is a much needed tool in todays strongly interdependent and complex world where over-simplistic assumptions frequently lead to poor decisions.

On one hand UbiSOM provides the daily correlation between products and can be made self-adjustable to continuously changing streams of data (i.e., both collaborative learning [Silva and Marques, 2010b] and detecting concept drift [Silva *et al.*, 2012]). On the other hand Ramex-Forum graphs shows sequences of the more representative events and can be easily used to model the dynamic of the occurrences. So, we believe that these complementary tools, one more static and the other more dynamic, can intrinsically guarantee realistic modeling on different scenarios and provide a major breakthrough in decision support systems.

References

[Agrawal and Srikant, 1995] R. Agrawal and R. Srikant. Mining sequential patterns. In *Proceedings 11th International Conference Data Engineering*, pages 3–14. IEEE Press, 1995.

[Kohonen, 1982] Teuvo Kohonen. Self-organized formation of topologically correct feature maps. *Biological cybernetics*, 43(1):59–69, 1982.

[Marques and Cavique, 2013] Nuno Marques and Luis Cavique. Sequential pattern mining of price interactions. In *Proceedings 16th Portuguese Conference on Artificial Intelligence (EPIA)*, 2013.

[Prim, 1957] Robert Clay Prim. Shortest connection networks and some generalizations. *Bell system technical journal*, 36(6):1389–1401, 1957.

[Silva and Marques, 2010a] B. Silva and N. Marques. Feature clustering with self-organizing maps and an application to financial time series portfolio selection. In *International Conference on Neural Computation*, 2010.

[Silva and Marques, 2010b] Bruno Silva and Nuno C. Marques. Ubiquitous data-mining with self-organizing maps. In *Proceedings of the Ubiquitous Data Mining Workshop, ECAI 2010*, 2010.

[Silva *et al.*, 2012] Bruno Silva, Nuno Marques, and Gisele Panosso. Applying neural networks for concept drift detection in financial markets. In *Workshop on Ubiquitous Data Mining*, page 43, 2012.

[Ultsch and Herrmann, 2005] A. Ultsch and L. Herrmann. The architecture of emergent self-organizing maps to reduce projection errors. In *Proceedings of the European Symposium on Artificial Neural Networks (ESANN 2005)*, pages 1–6. Verleysen M. (Eds), 2005.