BEHAVIOR ANALYSIS OF FINANCIAL MARKETS USING MARKOV MODELS, GENETIC ALGORITHMS, AND NEURAL NETWORKS

Logan Kelsch & Dr. Jing

I. Abstract

This research aims to identify anomalies indicative of underlying behavior within financial markets through a unique integration of several AI techniques. The methodology includes robust data collection, extensive variable construction (feature engineering), dimensionality reduction, normalization methods, and the use of a Genetic Algorithm (GA) applied to Markov Models (MM) and Neural Networks (NN). The anticipated outcomes will be a collection of critical insights into predictable behaviors or state-action representations in financial markets.

II. Introduction

Financial markets exhibit complex and often unpredictable behavior. With the newest era of Artificial and Computational Intelligence, recognizing and quantifying market behavior stands as one of the most unique challenges with one of the most recognized rewards. This research proposes a systematic exploration of data using machine learning & searching methods and statistical modeling techniques programmed in Python. The project will focus on developing an automated flexible search through the feature-space of collected data for areas that can converge to form truly predictive models. Specifically, the research will address the following key questions:

1. What data collection and feature engineering strategies maximize the informational richness relevant to market anomalies?
2. How can Markov Models and neural networks effectively represent underlying market states and actions?
3. In what manner can genetic algorithms refine these models to identify and exploit predictable anomalies?

III. Methods

This research is primarily quantitative and computational, organized into the following chronological phases:

1. **Data Acquisition & Feature Engineering:** Collect high-resolution market data (Open, High, Close, Volume, Time) across multiple indices and timeframes; develop a flexible pipeline for extracting and aggregating features.
2. **Prototype Model Implementations:** Build initial proof-of-concept instances of neural networks, transition matrices, and Markov Decision Processes (MDPs).
3. **Target-Space Development:** Design methods for defining target events, including price-action segmentation and event initiation/completion theory.
4. **Genetic Algorithm Prototyping:** Implement early GA versions with simple, pattern-based gene representations to validate the efficiency of a flexible search in complex space.

5. **Gene Population Management:** Develop serialization and deserialization routines for gene populations, and enable custom population initialization based on stored gene information.
6. **Gene Structure Definition:** Define comprehensive gene structures supporting mutative variability in neural network and Markov-model parameters.
7. **System Optimization & Integration:** Optimize performance of all components and integrate into a fully automated, end-to-end search algorithm.
8. **Large-Scale Feature & Target Search:** Execute extensive searches over feature and target spaces to identify models exhibiting robust, generalized predictive behavior.
9. **Walk-Forward Evaluation:** Conduct walk-forward backtesting on unseen data to assess model stability and out-of-sample performance.
10. **Results Compilation & Reporting:** Aggregate findings, analyze performance metrics, and prepare comprehensive reports detailing methodology, results, and conclusions.

IV. Technical Preliminaries

Min-Max Normalization

A collection of feature-space will be generated through the inter-relation of an observable variable θ over some length of time $\Delta\tau$. This will be measured using Min-Max normalization, which effectively scores the recent direction of movement of θ. Min-Max normalization will be applied as follows:

$$\widetilde{\theta}(\tau) \;=\; \frac{\theta(\tau) \;-\; \min\limits_{t\in[\tau-\Delta\tau,\,\tau]} \theta(t)}{\max\limits_{t\in[\tau-\Delta\tau,\,\tau]} \theta(t) \;-\; \min\limits_{t\in[\tau-\Delta\tau,\,\tau]} \theta(t)} \quad [1]$$

Dimensionality Reduction

Principal Component Analysis (PCA) will be used occasionally to reduce feature dimensionality and excessive computation, capturing the most significant variance and streamlining subsequent analyses.

Hawkes Process (Discrete-Time Approximation)

A Hawkes process is like the ripples in a pond after tossing a stone: each ripple (event) generates smaller ripples that gradually fade away. It is a self-exciting point process whose conditional intensity $\lambda(\tau)$ depends on the history of past events. In this study, a collection of feature-space will be generated by approximating the continuous-time Hawkes process by a discrete-time recursion updated at uniform intervals $\Delta\tau$. At each time step τ, $\lambda(\theta_\tau)$ is given by

$$\lambda(\theta_\tau) \;=\; e^{-\kappa\,\Delta\tau}\,\lambda\big(\theta_{\tau-\Delta\tau}\big) \;+\; \theta_\tau \quad [2,3]$$

where:

- $\lambda(\theta_\tau)$ is the intensity at time τ.
- $e^{-\kappa\,\Delta\tau}$ is the decay factor, controlling how much of the previous intensity is retained.
- κ is the inter-rater reliability over $\Delta\tau$ (rate of agreement).
- θ_τ denotes the significance of events (exogenous input) observed in the interval $(\tau-\Delta\tau,\ \tau]$.

For stability, we require

$$0 < \Delta\tau \text{ , and } 0 \;\leq \kappa \leq\; 1$$

ensuring excitation cannot accumulate without bound.

Genetic Algorithm (GA)

Genetic algorithms will be used to explore the feature-space for areas that contain indications of predictable market behavior. The steps for this algorithm include:

- **Gene Construction**: The architecture of the genes will take several forms, but ultimately stand as methods (MMs, NNs, pattern formation) of observing and evaluating some collection of exogenous information.
- **Target and Gene Population Initialization**: A set of solutions for each instance of time will be calculated from several practical approaches, all measuring the displacement of some segment of price action into the future. An initial population of genes is also generated at random or from desired specifications.
- **Fitness Evaluation**: Each gene's response to exogenous information across various types of markets will act as a method of predicting market behavior and will be scored on consistency, frequency, and efficiency using metrics such as Pearson's r^2, Martin Ratio, Sharpe Ratio and others.
- **Elite Selection**: The best genes will be exempt from not being selected in the parent selection step to prevent extinction and ensure greediness in the search.
- **Parent Selection**: A random set of genes that perform above a survival threshold are selected for reproduction.
- **Gene Reproduction**: Of the selected parent genes, a set of random combinations of their sources of exogenous information or interpretation processes are generated as the child genes for the following generation.
- **Gene Mutation**: Because the feature-space being searched contains essentially incalculable complexity, a randomization step is enforced. This provides a small probability of interpretation methods or exogenous sources of a gene to shift to something entirely different.

Once the algorithm begins finding working predictive models, a model correlation transformation will have to be made to the target-space in the Target Initialization step to enforce subsequent searching to be done in novel areas.

Neural Networks (NN)

Neural networks will be implemented using TensorFlow and Keras with CUDA acceleration. In this study, they serve as genes in the genetic-algorithm search. Variability will be introduced across the following dimensions:

- **Architecture depth:** number of layers and neurons per layer
- **Optimizer choice:** Adam, SGD, etc.
- **Learning rate:** between 10^{-4} and 10^{-2}
- **Output type:** regression or classification, set by Fitness Evaluation method

Potentially, A difference in Neural Network type may be considered (FF, LSTM, CNN, other). Each network will be trained via backpropagation with early stopping.

Markov Models (MM)

Markov models—including Markov Decision Processes (MDPs) and transition matrices—will be implemented using established libraries (e.g. `hmmlearn`, `pomegranate`). In this study, these models will be implemented as genes in the Genetic Algorithm Search. The gene variability will likely include:

- **State Space Size:** generating states based off of distribution of exogenous information

- **Transition probability initialization:** Random uniform, Dirichlet draws
- **Emission distribution type:** Gaussian, categorical, or mixed
- **Reward function:** Parameters and discount factor γ
- **Solver algorithm choice:** value iteration vs. policy iteration

Parameters will be trained or inferred via dynamic programming. Each model's fitness will be evaluated on its ability to reconstruct observed market state sequences or maximize expected return under simulated trading scenarios.

V. Predicted Results

After gathering a collection of various model types with custom parameters and exogenous sources, we can expect to have gained several insights:

- **Identification of Discernible Behavior & Robust Anomalies:** We expect to decrypt a concise set of market behaviors in one way or another by models generated by the Genetic Algorithm. This could show itself as consistent state dynamics, differentiation of predictable and non-predictable market environments, or much more complex data interpretations.
- **Model Type Effectiveness:** Comparative analysis using the Genetic Algorithm will reveal which families of models (MDP vs. NN) perform best overall, as well as uncover the differences in performance across different areas in the target-space.
- **Optimal Model Parameters:** The Genetic algorithm results will converge on preferred parameter settings—such as optimal state counts, network depths, and decay rates—demonstrating efficient evolutionary searches that balance complexity and predictive performance.

VI. Significance

The integration of genetic algorithms with Markovian and neural network models expects to turn a complex and extensive market data search into actionable insights. By automating the search for behavioral anomalies, this research will uncovers patterns that traditional methods miss entirely. Rigorous walk-forward evaluations will ensure that our findings are not merely academic curiosities, but hold up under real-world, out-of-sample testing.

Beyond improving short-term forecasting and anomaly detection, the techniques developed here will constitute a versatile toolkit for any domain where self-exciting dynamics and latent states play a role—from algorithmic trading to epidemiology and beyond. Ultimately, this project will deliver a polished, end-to-end system capable of identifying and exploiting predictable market behaviors, providing both theoretical contributions and practical value to researchers and investors alike.

VII. Budget

For this project, all programming and code execution will be done on my own computer. I do believe that the RAM, CPU, and GPU of my computer will be sufficient for this project.

References

[1] Jiawei Han, Micheline Kamber, and Jian Pei, *Data Mining: Concepts and Techniques*, 3rd ed., Morgan Kaufmann, 2011.

[2] Alan G. Hawkes, “Spectra of some self-exciting and mutually exciting point processes,” *Biometrika*, vol. 58, no. 1, pp. 83–90, 1971.

[3] Lorick Huang and Mahmoud Khabou, “Convergence of the Discrete-Time Compound Hawkes Process with Exponential or Erlang Kernel,” arXiv:2106.13459 [math.PR], 2021.