"An Exploratory Study on the Complexity and Machine Learning Predictability of Stock Market Data" – Sebastian Raubitzek and Thomas Neubauer

Summary by Nossa Iyamu

Objective

Discover trends and correlations between predictability and complexity.

Data

Dow Jones Industrial Average, NASDAQ, M1-money supply until 31 December 2019.

Model used

Stochastic gradient descent linear regression, Lasso regression, XGBoost tree regression.

Results

Increase in complexity $\Rightarrow$ decrease in predictability.

Methodology

- Each year treated separately.
- Complexity measured with Fisher's information, Shanon's Entropy, Approximate Entropy, Sample Entropy, Fractal dimension, Hurst exponent, error of Hurst exponent.
- 100 predictions problem for each dataset differing in their memory of previous values.
- Split in train test: 0.8 / 0.2.
- Regression analysis with Machine learning.

Previous Research

Hurst exponent higher than 0.5 indicates a more predictable stock market, and $H = 0.5$ indicates a random motion (Brownian motion).

Machine Learning Approach

- Different models to have different regression approaches.
- Optimized with RandomizeSearchCV.
- Error measured with RMSE, R^2-score for test sets, cross-validation for train sets.

Complexity Analysis Approach

- Separating M1-trended and M1-detrended.
- Measure complexity with every metric mentioned above.
- Find correlation between predictability with ML and complexity.

Results

- Hurst exponent decreased in later years, now it's ~ 0.5.
- Shanon's entropy increases with years.
- Fisher info increases but fluctuates a lot.
- Stagnation of Sample entropy and App Entropy.
- Lasso performed better than the other models.
- SGD performed the worse, and SGD take lower number of input data for a max performance $\Rightarrow$ witness a memory in stock market
- M1-detrended significant on NASDAQ, improves predictability for some regions but worsens for others.
- R^2 score highly correlated with Sample Entropy for Dow Jones, and with Approximate Entropy for NASDAQ.
- Authors say that these results hold for every dataset.

Conclusion

- Stock market more random and unpredictable for later years than for earlier years $\Rightarrow$ more similarities with a fractional Brownian motion.
- M1-detrended increases randomness for earlier years, for later years: no additional noise $\Rightarrow$ should be the topic of future studies (authors interpret this as varying money supply already influencing the data; maybe the variation is more important for later years).

- Did not predict future trends but performed regression analysis on dataset under study.
- High entropy $\Rightarrow$ Low predictability, vice versa. Thus, it's possible to find stock market periods with high predictability.
- Advice for future study: employ complexity measures and ideas from Chaos theory to improve machine learning approaches.
- Article [16] and [11] predicted periods with high Hurst exponent using neural networks.