Defending Mathematics: An Analysis of Backtest Overfitting in Financial Trading Algorithms

Stephanie Ger[1], Alex Sim[2], John Wu[2]
[1]Boston College, [2]Lawrence Berkeley National Laboratory,

ABSTRACT

Backtests for trading algorithms on historical data are often used to gain an understanding of how the algorithm will perform in the future. Oftentimes, investigators will fail to report the number of algorithms that were tested or the out of sample Sharpe ratio. Without those statistics, it is difficult to determine how much predictive power the algorithm has. A website was build to display how easy it can be to overfit data by comparing how an algorithm optimized to a data set performs on that data to its performance on a new data set. In addition, though a formula for minimal backtest length exists, it calculates the number of independent trials that can be run on a given data set. In order to determine the number of independent trials exist in a given sample of trials, applications of information theoretic techniques such as entropy and multi-information functions have been studied.

BACKGROUND INFO

- Algorithmic models are usually based on some observation of stock market behavior
- Algorithms with a high Sharpe Ratio is more likely to sell, so companies will not report trials that performed poorly
- The the out of sample (OOS) Sharpe Ratio or number of trials
- The role that backtest overfitting plays in algorithm development has attracted the attention of regulatory agencies such as the Securities and Exchanges

RESEARCH QUESTION

The goal of this project was to increase awareness about the dangers of overfitting data when conducting backtests and then to determine how many trials can be run on a data set given its length.

METHODS AND MATERIALS

The code that is used to optimize the algorithm for a given data set calculates the Sharpe ratio for each of the four parameters:

- Holding Period
- Side
- Stop Loss
- Entry Day

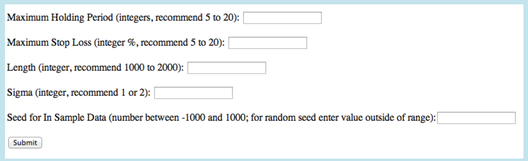
and outputs a new iteration of the algorithm only if the new Sharpe ratio exceeds the previous Sharpe ratios.

INDEPENDENCE OF TRIALS

- Minimal backtest length for independent trials known
- Cannot assume trials are independent
- The use of correlation matrices leads to data overfitting
- Alternatives to correlation matrices include the use of nonparametric dependence estimators based entropy measures

WEBSITE

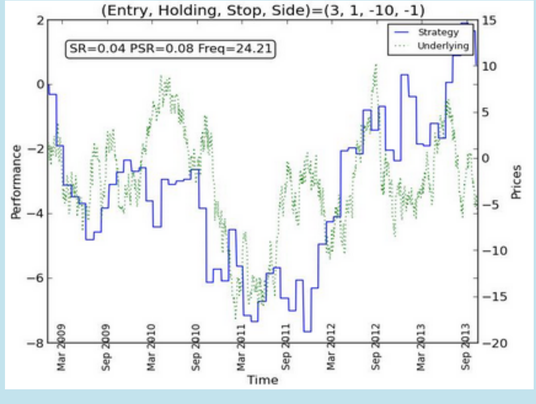
The website displays the differences between the algorithm's performance on the OOS versus IS data.



Input form for website

Once the form is submitted, the Python code runs, and outputs a video that shows how the algorithm is optimized and then how it performs on the OOS data.

Result of Python code



Result of Python code

CONCLUSION

- it is clear that asides from the disparity in the Sharpe ratios, that the algorithm performs differently in the two data samples
- in both images, the blue line represents how the algorithm performs on the data set, which is represented by the dotted green line
- In sample, the algorithm displays a positive trend with smaller spikes so it is generating profit
- Out of sample, the blue line is characterized by large spikes, which is indicative of random behavior
- the increased variability present in the out of sample data demonstrates the loss in predictive power that the algorithm suffers due to the overfitting
- ideas from entropy functions have provided insight into how to approach the problem of determining the number of independent trials.

ACKNOWLEDGMENTS

We thank Marcos Lopez de Prado for the code he provided. We also thank David H. Bailey, Jonathan M. Borwein and Qiji Jim Zhu for their work on backtest overfitting. This work was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internship (SULI) program.