HUDSON AND THAMES

Advanced Topics in Financial Machine Learning

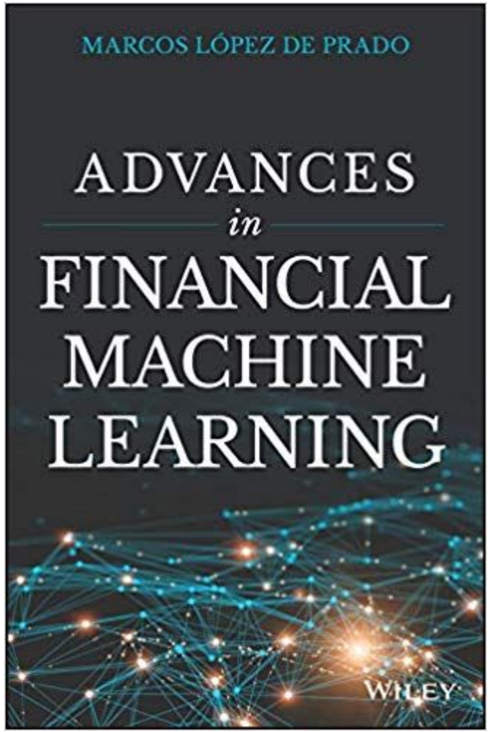
HUDSON AND THAMES

Table of Contents

We believe that the scientific method is the best way to approach investment management.

This presentation highlights capability and use cases for machine learning in Finance. In particular it covers techniques from the body of literature that our team is most familiar with.

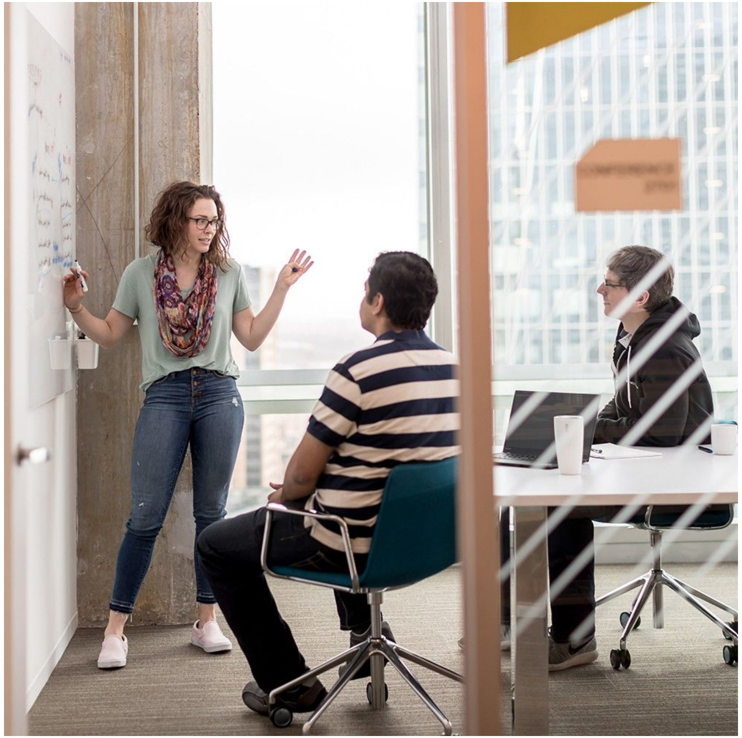
Advances in Financial Machine Learning



www.quantresearch.org

HUDSON AND THAMES

Machine Learning in Finance



1. Avoid research through backtesting.
2. Improving the statistical properties of your underlying data.
3. Filtering events which are more statistically predictable.
4. Labeling Techniques
5. Sample Weights
6. Feature Engineering
7. Market Microstructural Features
8. Feature Importance
9. Optimising trading rules without Backtesting
10. Cross Validation in Finance
11. **Sequentially Bootstrapped Ensembles**
12. **Filtering out False Positives** (Boost Sharpe Ratio)
13. Optimal Bet Sizing Strategies
14. **Portfolio Optimisation** that has been shown to outperform competitor algorithms, out-of-sample
15. **Detection of False Investment Strategies**

Statistical Properties - Profiling

Variables

CHEFTYOY Index
Numeric

Distinct count	79	Mean	0.39756
Unique (%)	12.2%	Minimum	-8.2
Missing (%)	81.0%	Maximum	7.8
Missing (n)	525	Zeros (%)	0.2%
Infinite (%)	0.0%		
Infinite (n)	0		

Toggle details

CHFADC Index
Numeric

Distinct count	124	Mean	107590
Unique (%)	19.1%	Minimum	39526
Missing (%)	81.0%	Maximum	203920
Missing (n)	525	Zeros (%)	0.0%
Infinite (%)	0.0%		
Infinite (n)	0		

Toggle details

~~CHFAGOVT Index~~
Highly correlated

This variable is highly correlated with `CHFAOFIN Index` and should be ignored for analysis

Correlation 0.92495

~~CHFANFS Index~~
Highly correlated

This variable is highly correlated with `CHFADC Index` and should be ignored for analysis

Correlation 0.9955

BBDXY Index
Numeric

Distinct count	2492	Mean	1085.5
Unique (%)	92.3%	Minimum	912.58
Missing (%)	1.6%	Maximum	1277.5
Missing (n)	44	Zeros (%)	0.0%
Infinite (%)	0.0%		
Infinite (n)	0		

Toggle details

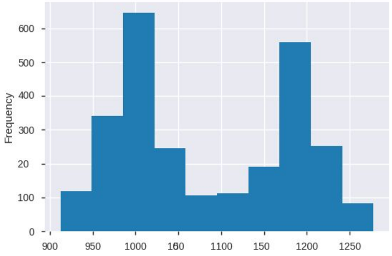
Statistics | Histogram | Common Values | Extreme Values

Quantile statistics

Minimum	912.58
5-th percentile	954.01
Q1	1002.8
Median	1051.6
Q3	1183.7
95-th percentile	1230.5
Maximum	1277.5
Range	364.95
Interquartile range	180.94

Descriptive statistics

Standard deviation	98.394
Coef of variation	0.090647
Kurtosis	-1.4942
Mean	1085.5
MAD	91.318
Skewness	0.15098
Sum	2881900
Variance	9681.3
Memory size	21.2 KiB



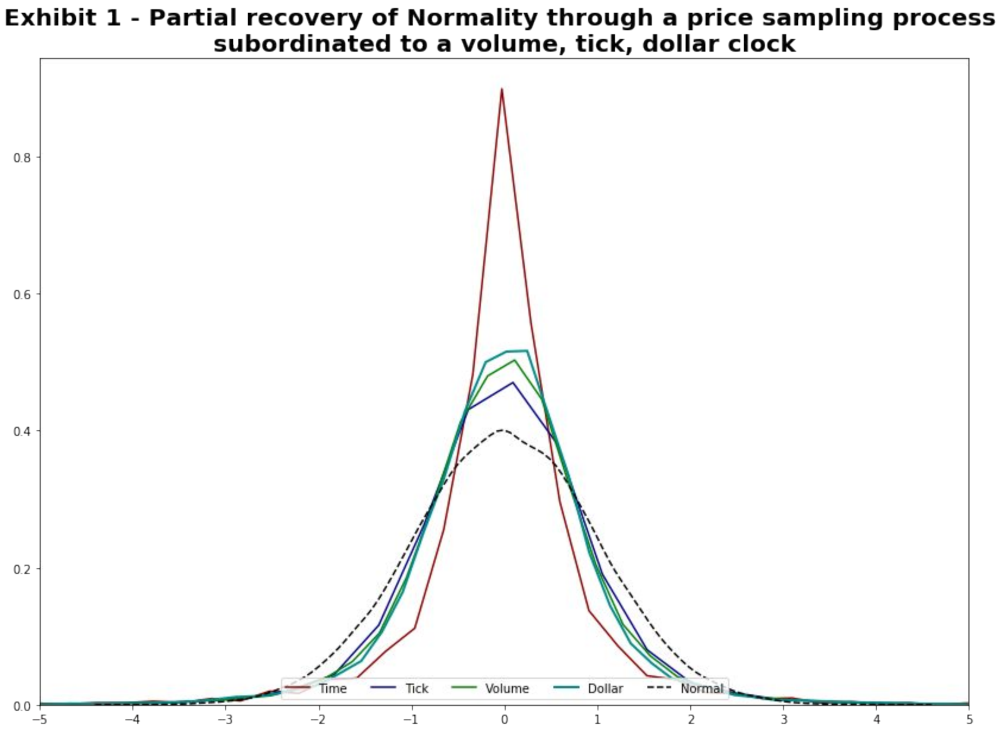
Minimum 5 values

Value	Count	Frequency (%)
912.58	1	0.0%
914.4	1	0.0%
915.02	1	0.0%
915.52	1	0.0%
916.48	1	0.0%

Maximum 5 values

Value	Count	Frequency (%)
1273.52	1	0.0%
1274.25	1	0.0%
1274.84	1	0.0%
1276.72	1	0.0%
1277.53	1	0.0%

Better Sampling Techniques



> Chronological Sampling (fixed time interval sampling)
> New Financial Data Structures

Standard Bars:

- Tick Bars
- Volume Bars
- Dollar Bars

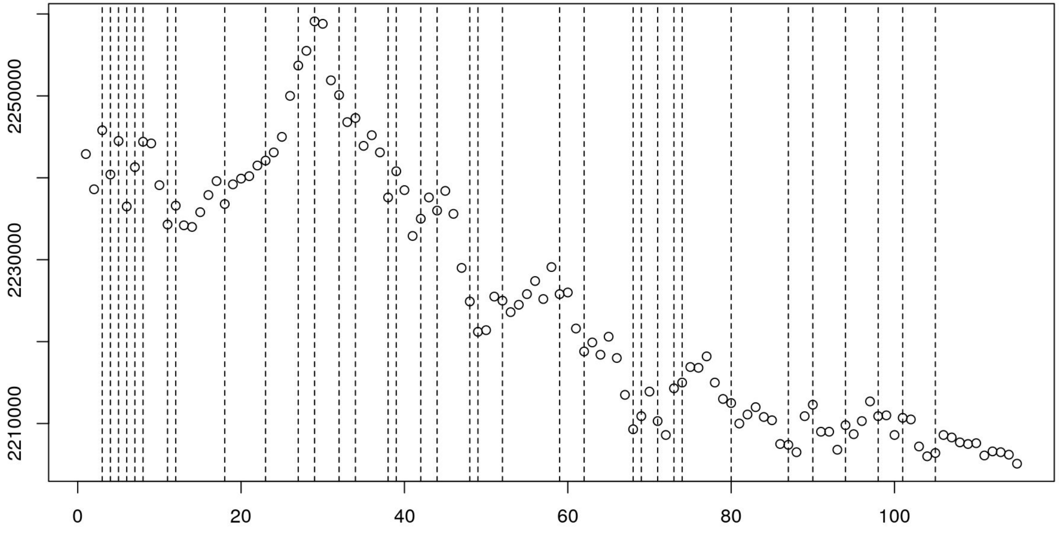
Information Driven Bars

- Imbalance Bars
- Run Bars

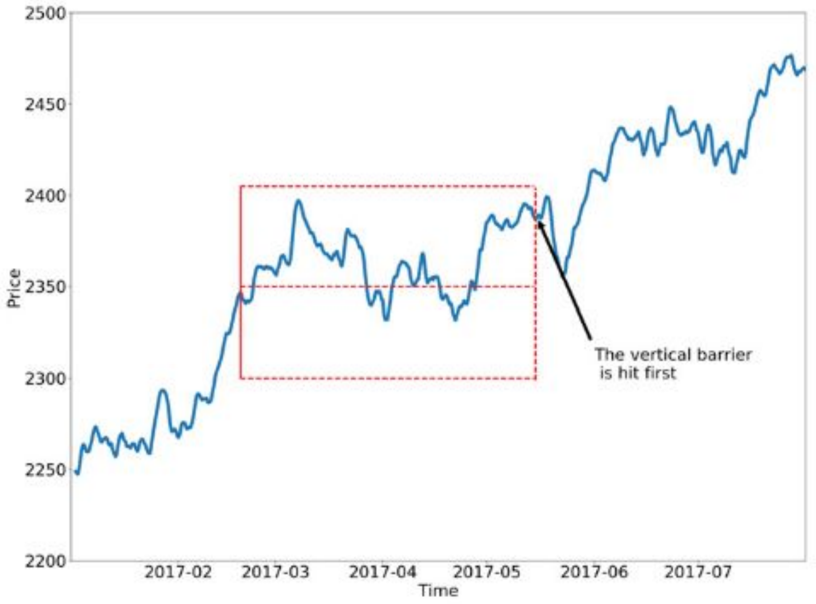
Filtering Events

Strategy Triggers:

- Momentum
- Weather (Energy)
- Structural Breaks
- Order Imbalance



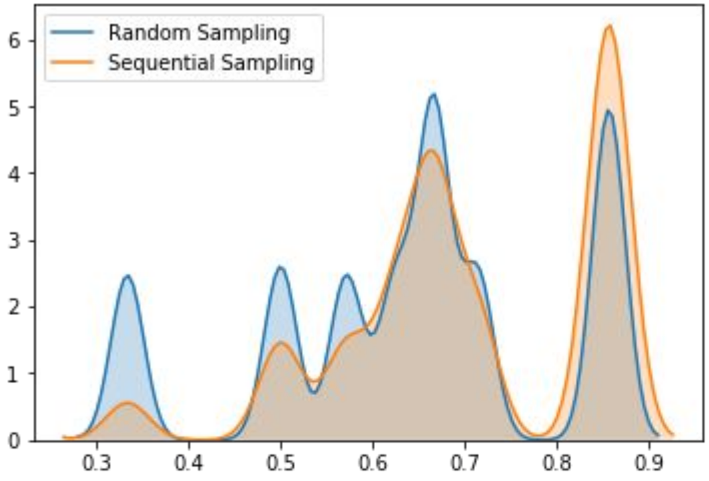
Financial Labeling Techniques: Triple-Barrier



(b)

- The Triple Barrier Method labels an observation according to the first barrier touched out of three barriers.
  - Two horizontal barriers are defined by profit-taking and stop-loss limits, which are a dynamic function of estimated volatility (whether realized or implied).
  - A third, vertical barrier, is defined in terms of number of bars elapsed since the position was taken (an expiration limit).
- The barrier that is touched first by the price path determines the label:
  - Upper horizontal barrier: Label 1.
  - Lower horizontal barrier: Label -1.
  - Vertical barrier: Label 0.

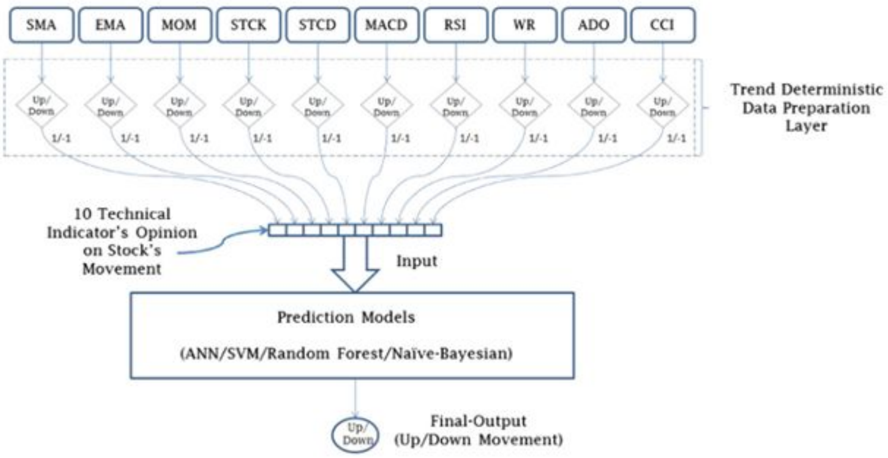
Sample Weights



- In financial machine learning, samples are not independent
- Samples suffer from a low average uniqueness.
- Can make use of sampling techniques to boost model performance.
- See our implementation of Sequentially Bootstrapped Ensembles.

Feature Engineering

Trend Deterministic Data Preparation



Fractional Differentiation

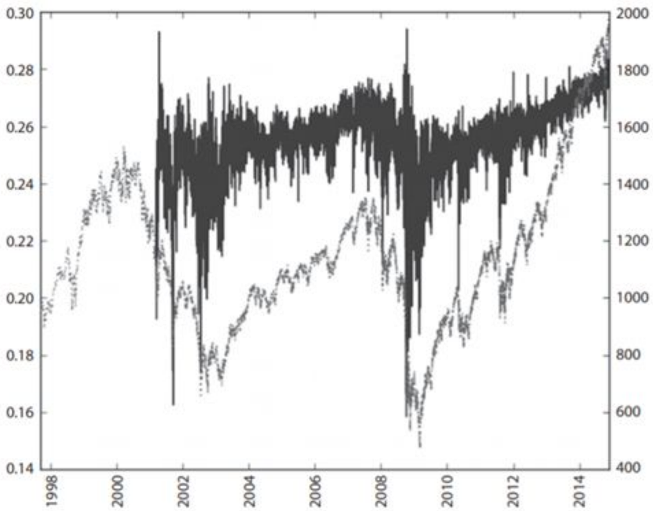
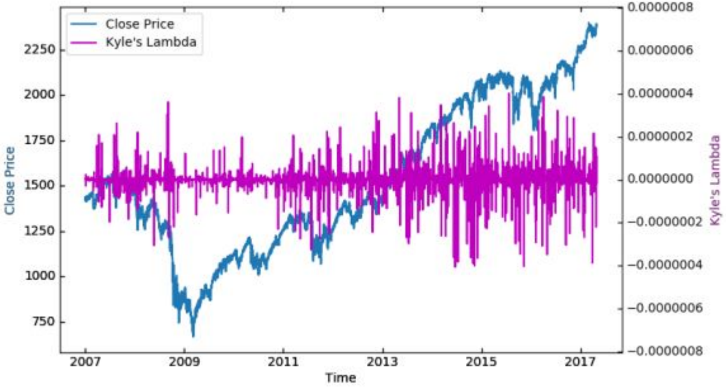


FIGURE 5.4 Fractional differentiation after controlling for weight loss with a fixed-width window

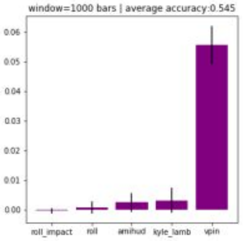
Market Microstructural Features

Microstructural datasets include primary information about the auctioning process, like order cancellations, double auction book, queues, partial fills, aggressor side, corrections, replacements, etc.

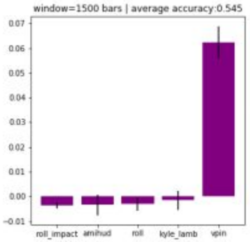
That makes microstructural data one of the most important ingredients for building predictive ML features.



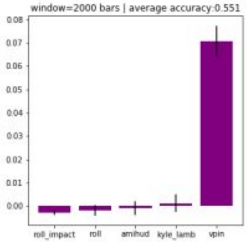
1. Roll Measure
2. Roll Impact
3. Kyle's Lambda
4. Amihud's Lambda
5. Hasbrouck's Lambda
6. VPIN



1000 bars

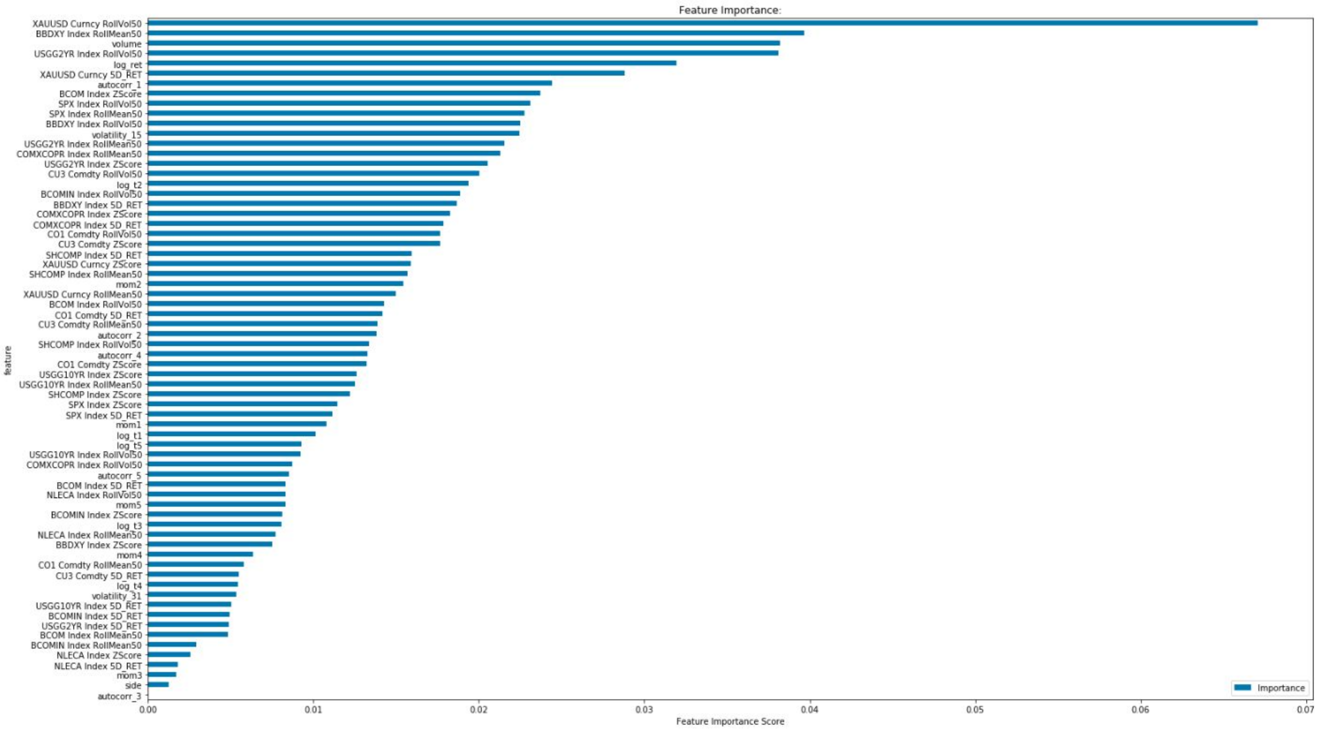


1500 bars



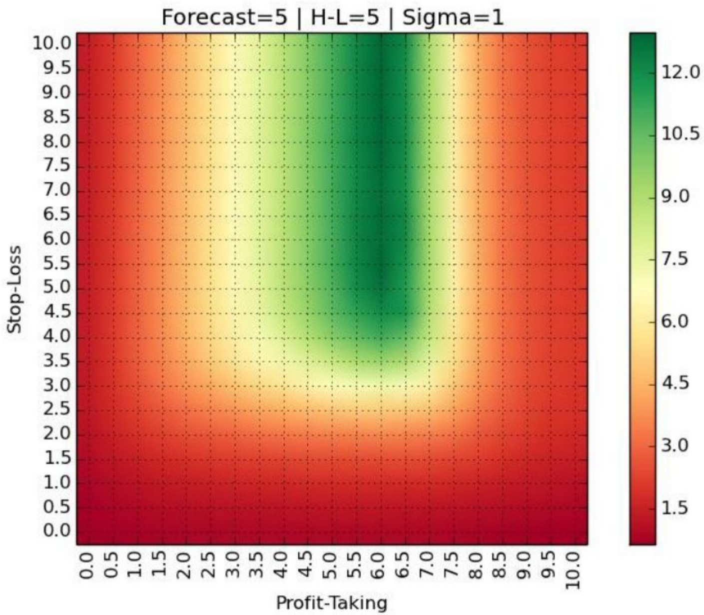
2000 bars

Feature Importance



- Avoid research through backtesting - multiple testing - increases the probability of making a false discovery.
- Better to focus on feature importance.
  - Engineer useful features
  - Drop those that contribute to noise
- Importance algorithms:
  - Mean decrease impurity (MDI)
  - Mean decrease accuracy (MDA)
  - Single feature importance (SFI)

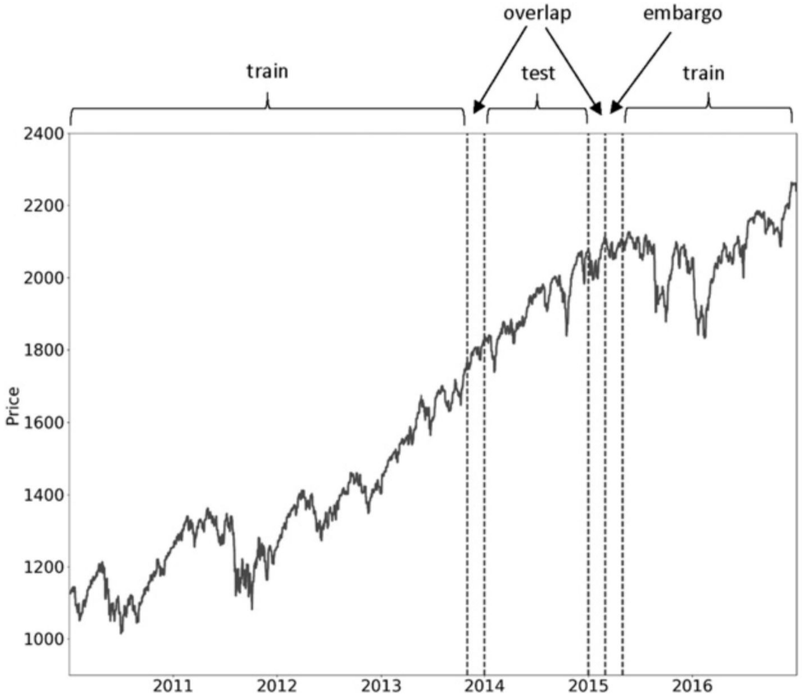
Optimal Trading Rules without Backtesting



Calibrating a trading rule using a historical simulation contributes to backtest overfitting, which in turn leads to underperformance.

Can use synthetic data generated using stochastic processes such as the Ornstein-Uhlenbeck process to help determine optimal parameters, without overfitting.

Purged & Embargoed K-Fold CV



Standard CV fails in a finance setting due to:

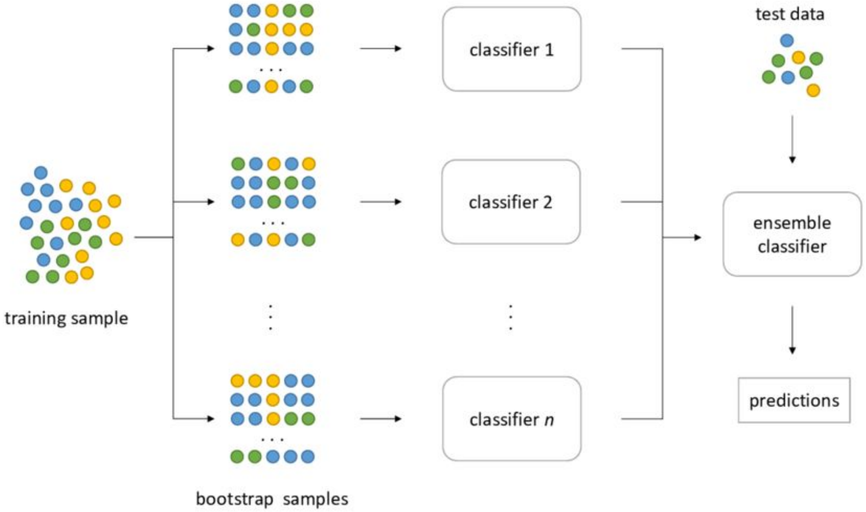
- Observations can't be assumed to be drawn from an IID process.
- Leads to multiple testing and selection bias.
- Leakage takes place when training set contains info that appears in testing set. This happens as a result of shuffling, and overlapping samples.

Resolved using Purged and Embargoed K-Fold CV

Ensembles

An ensemble method is a method that combines a set of weak learners, all based on the same learning algorithm, in order to create a (stronger) learner that performs better than any of the individual ones. Ensemble methods help reduce bias and/or variance.

1. Sequentially Bootstrapped Ensembles (H&T Implementation)
2. Voting Classifiers
3. Bagging
4. Stacking



Filtering False Positives: Boost Sharpe Ratio

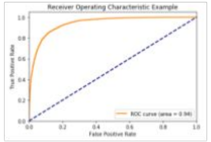
Recommend two separate models.

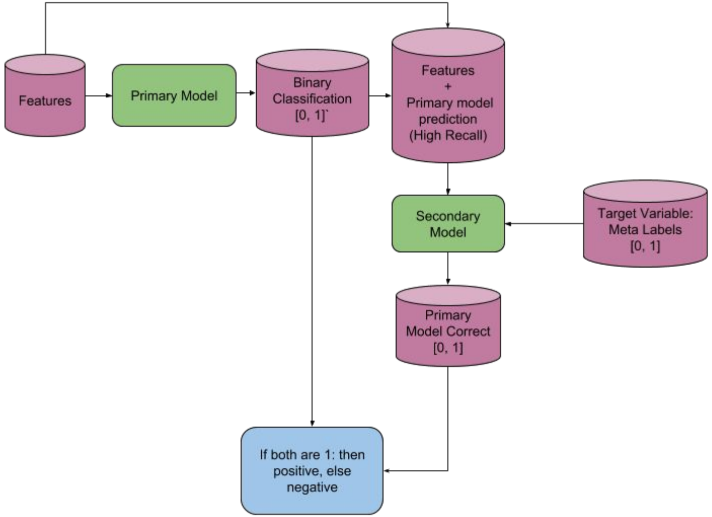
1. Side of the position (alpha model)
2. Size of the position (risk management)

Meta-Labeling

- Takes the side from the primary model (long or short).
- Train a ML model to determine if we should trade on the signal or not.
  - Train Random Forest
  - Use Cross-validation and Grid Search to find the optimal hyperparameters.
- Map confidence level to position size
  - Add bet sizing algorithm

Meta Labeling

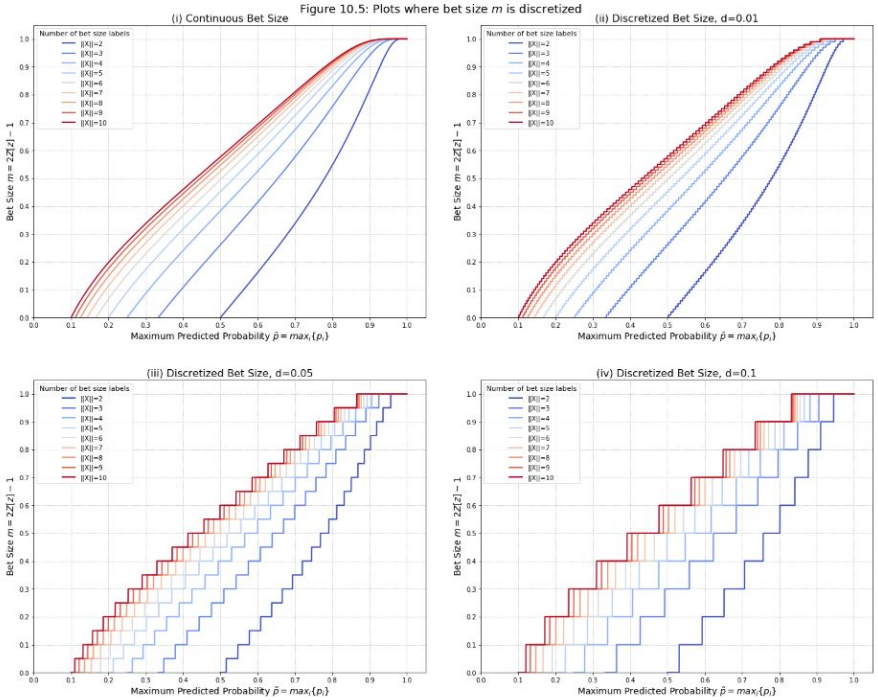




Optimal Bet Sizing

Assuming a machine learning algorithm has predicted a series of investment positions, one can use the probabilities of each of these predictions to derive the size of that specific bet.

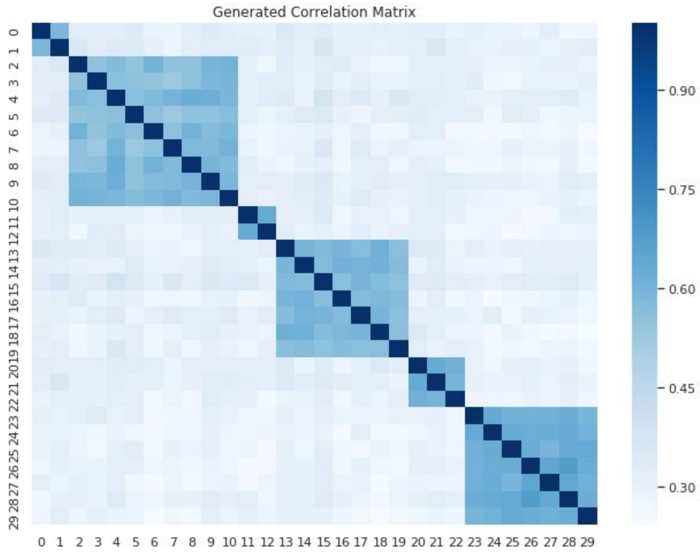
We have a number of bet sizing algorithms.



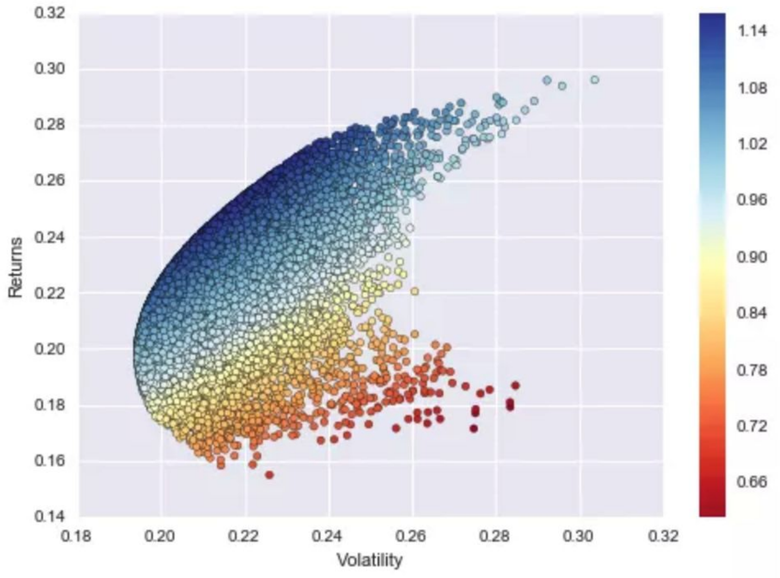
Detection of False Investment Strategies

3rd Law of Backtesting: "Every backtest result must be reported in conjunction with all the trials involved in its production. Absent that information, it is impossible to assess the backtest's 'false discovery' probability"

- Probabilistic Sharpe Ratio
- Deflated Sharpe Ratio
- Optimal Number of Clusters (Unsupervised Learning Algorithm)



Portfolio Optimization: Mean Variance



Searching for the efficient frontier (Python for Finance, 2017)

- Mathematical framework for assembling a portfolio of assets such that the expected return is maximized for a given level of risk.
- Maximise returns
- Reduce variance of returns
- Harry Markowitz (1952)

New Optimization: Hierarchical Risk Parity

Building Diversified Portfolios that Outperform Out of Sample

Marcos López de Prado

Marcos López de Prado
is a senior managing director at Guggenheim Partners in New York, NY, and a research fellow at the Lawrence Berkeley National Laboratory in Berkeley, CA.
lopezdeprado@lbl.gov

Portfolio construction is perhaps the most recurrent financial problem. On a daily basis, investment managers must build portfolios that incorporate their views and forecasts on risks and returns. Before Markowitz earned his Ph.D. in 1954, he left academia to work for the RAND Corporation, where he developed the Critical Line Algorithm (CLA), a quadratic optimization procedure specifically designed for inequality-constrained portfolio optimization problems. This algorithm is notable in that it guarantees finding the exact solution after a known number of iterations—and it ingeniously circumvents the Karush-Kuhn-Tucker conditions (Kuhn and Tucker [1952]). A description and open-source implementation of this algorithm can be found in Bailey and López de Prado [2013]. Surprisingly, most financial practitioners still seem unaware of CLA, as they often rely on generic-purpose quadratic programming methods that do not guarantee the correct solution or a stopping time.

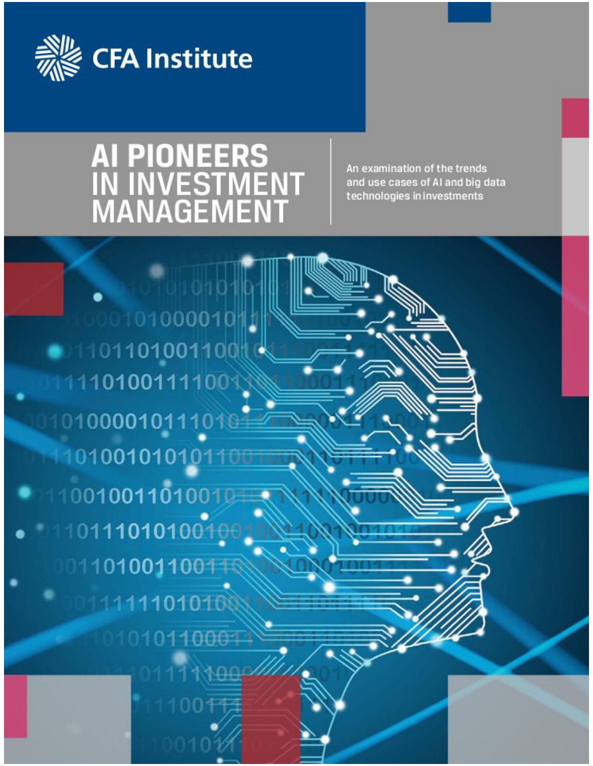
Despite of the brilliance of Markowitz's theory, CLA solutions are somewhat unreliable because of a number of practical problems. A major caveat is that small deviations in the forecasted returns cause CLA to produce very different portfolios (Michaud [1998]). Given that returns can rarely be forecasted with sufficient accuracy, many authors have opted to drop them altogether and focus on the covariance matrix. This has led to risk-based asset allocation approaches, of which "risk parity" is a prominent example (Jurczenko [2015]). Dropping the forecasts on returns improves the instability issues; however, it does not prevent them, because quadratic programming methods require the inversion of a positive-definite covariance matrix (all eigenvalues must be positive). This inversion is prone to large errors when the covariance matrix is numerically ill-conditioned—that is, it has a high condition number (Bailey and López de Prado [2012]).

MARKOWITZ'S CURSE

The condition number of a covariance, correlation (or normal, thus diagonalizable) matrix is the absolute value of the ratio between its maximal and minimal (by moduli) eigenvalues. Exhibit 1 plots the sorted eigenvalues of several correlation matrices, where the condition number is the ratio between the first and last values of each line. This number is lowest for a diagonal correlation matrix, which is its own inverse. As we add correlated (multicollinear) investments, the condition number grows. At some point, the condition number is so high that numerical errors make the inverse matrix too unstable: A small change on any entry will lead to a very different inverse. This is *Markowitz's*

- HRP does not require the invertibility of the covariance matrix.
- In fact, HRP can compute a portfolio on an ill-degenerated or even a singular covariance matrix, an impossible feat for quadratic optimizers.
- Monte Carlo experiments show that HRP delivers lower out-of-sample variance than CLA, even though minimum-variance is CLA's optimization objective.
- HRP also produces less risky portfolios out-of-sample compared to traditional risk parity methods.

Further Resources



Case Studies

1. Enhancing Trading Strategy and Execution with Machine Learning: Man AHL.
2. Generating Signals for Quant Models with Machine Learning: New York Life Investments.
3. Refining Equity Trading Volume Prediction with Deep Learning: State Street Corporation.
4. Leveraging AI/Alternative Data Analysis in Sell-Side Research: Goldman Sachs
5. Dissecting Earnings Conference Calls with AI and Big Data: American Century.
6. AI and Big Data Assist in Debt Portfolio Management: China Life Asset Management and China Securities Credit Investment.
7. Applying AI and Big Data Technologies in the Filing and Processing of Insurance Claims and Assessing Corporate Risk: Ping An.
8. Sentiment Analysis: Bloomberg.
9. Building the Data Science Team: Schroders.
10. Special Focus: Enhancing the MPT Efficient Frontier with Machine Learning.
11. Special Focus: Using Intelligent Searches to Collect and Process Information.

HUDSON
AND THAMES

Further Resources

Academic Journal:

1. A Backtesting Protocol in the Era of Machine Learning
2. Neural Networks in Finance: *Design and Performance*
3. Enhancing Time-Series Momentum Strategies Using Deep Neural Networks
4. Time-Series Momentum: *A Monte Carlo Approach*
5. Extracting Signals from High-Frequency Trading with Digital Signal Processing Tools
6. Industry Return Predictability: *A Machine Learning Approach*
7. A Machine Learning Approach to Risk Factors: *A Case Study Using the Fama–French–Carhart Model*
8. Big Data in Portfolio Allocation: *A New Approach to Successful Portfolio Optimization*
9. A Practical Approach to Advanced Text Mining in Finance
10. Dynamic Replication and Hedging: *A Reinforcement Learning Approach*

GitHub

HUDSON
AND THAMES

Thank You