Final Presentation

2024.01.17

Contents

1부

- 2023-02 총정리
- 최근 Quantitative Finance 분야 연구 소개
- 학회 활동 및 수상내역 소개

2부

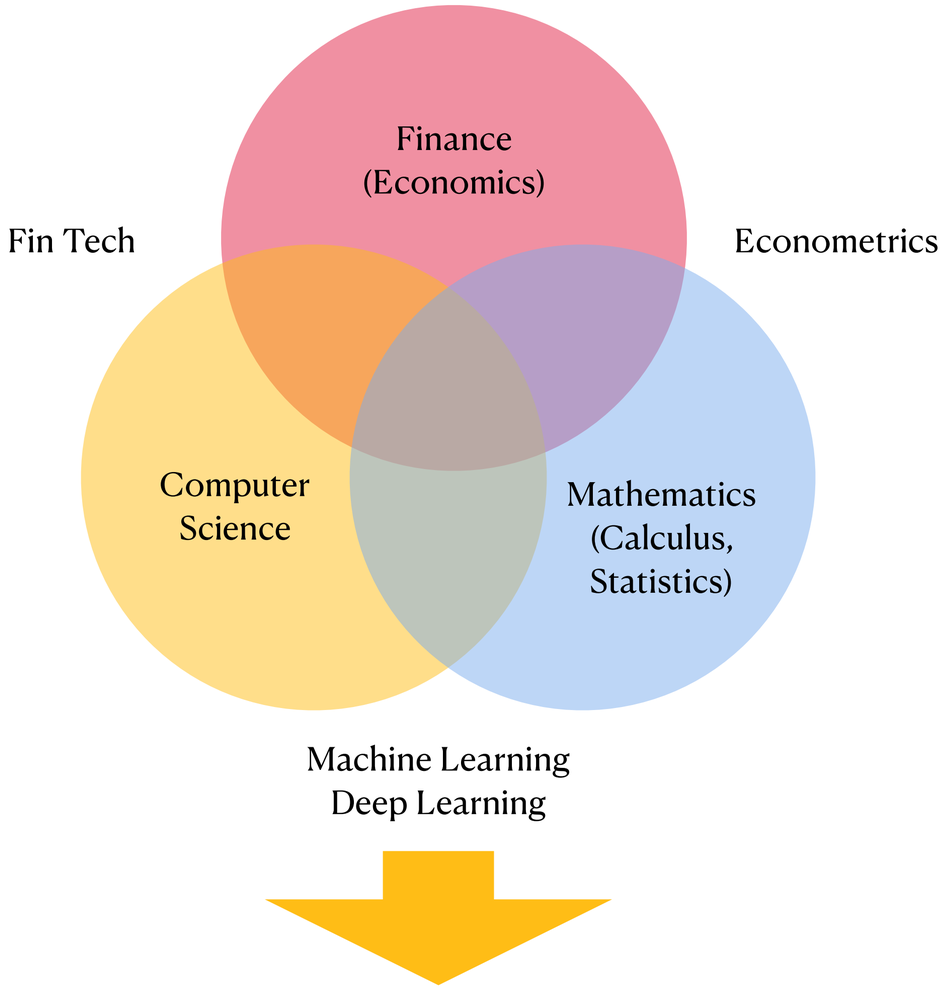
- 독립된 주제로써의 금융 머신러닝 (Financial Machine Learning) 연구를 위한 python library 제작
- Probabilistic Sharpe Ratio, Deflated Sharpe Ratio를 통한 Profit taking / Stopping loss threshold 지정
- Algorithmic Trading with Metha labeling strategies
- QA Chatbot

Quantitative Finance

2023-02

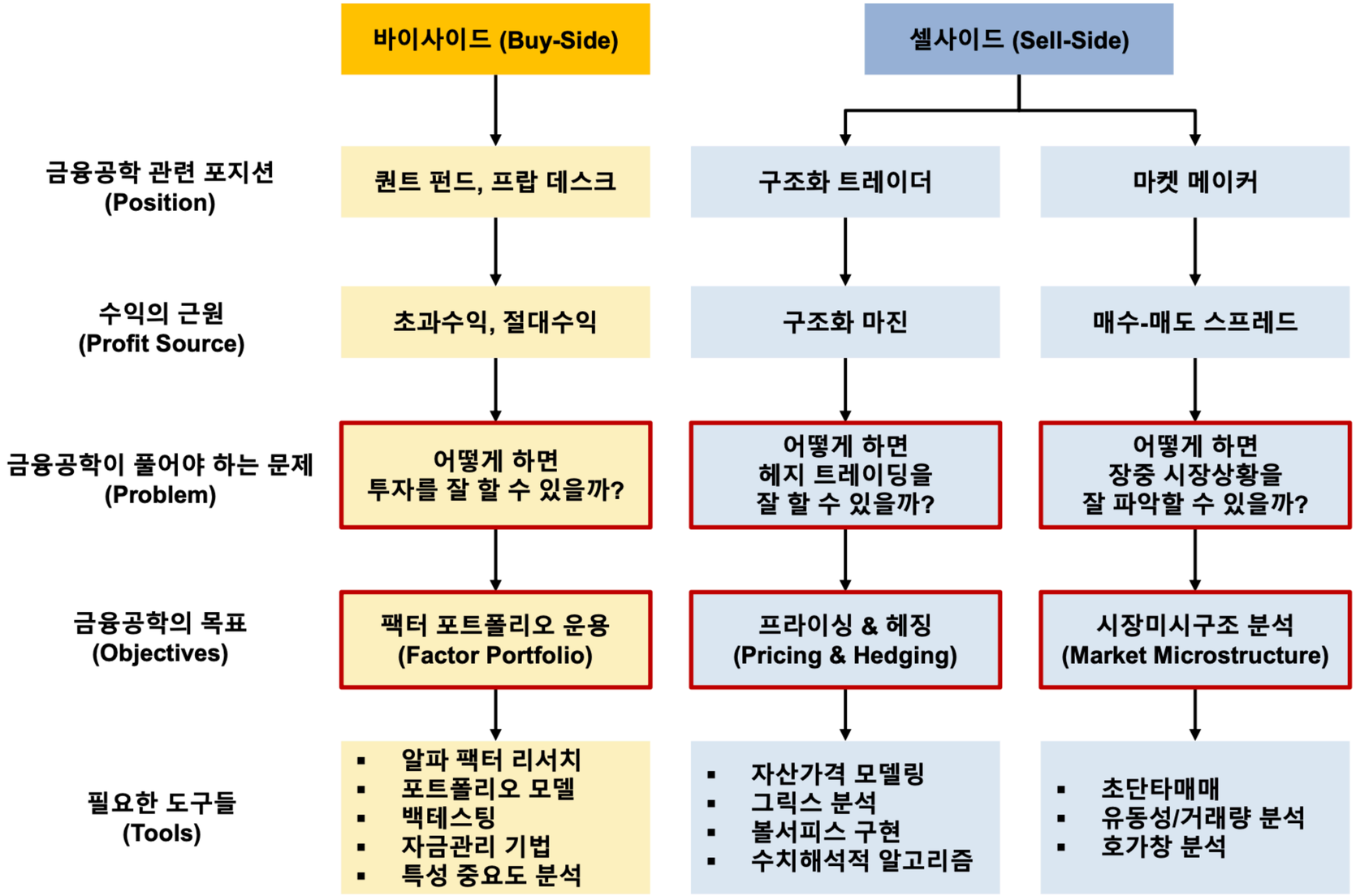
2024.01.17

What is Quant?

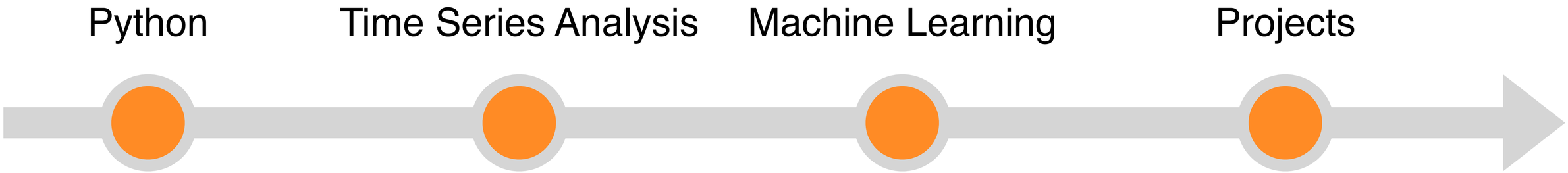


Quantitative Finance, Financial Engineering

What is Quant?



Basic Quantitative Finance



Python Basic
Numpy
Pandas
Matplotlib
Algebra, Statistics

Week 01-03

Basic Time Series
Linear Time Series
Heteroscedasticity
Signal denoising
Microstructure

Week 04-10

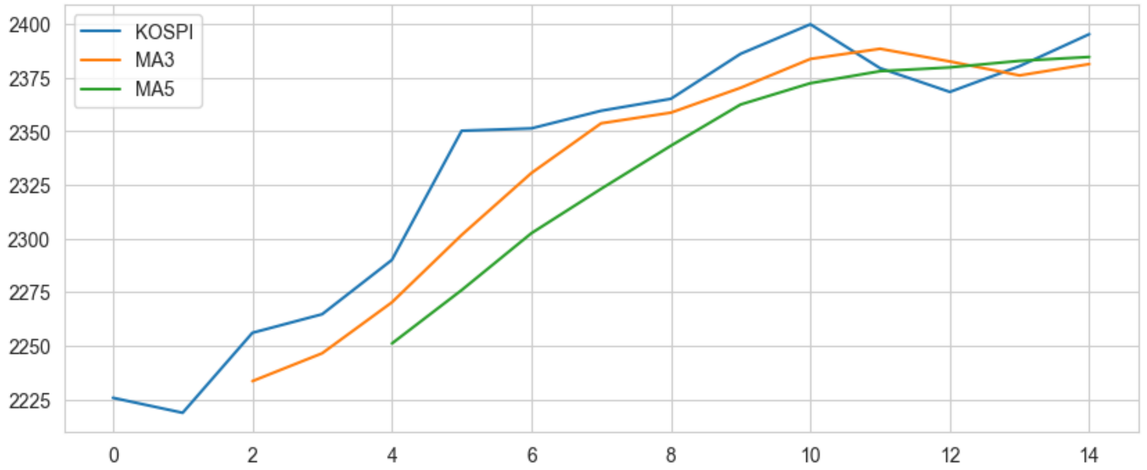
Linear Models
Regression
Classification
Dimension reduction
Clustering
Explainability

Week 11-16

Basic Programming Concepts

Week01

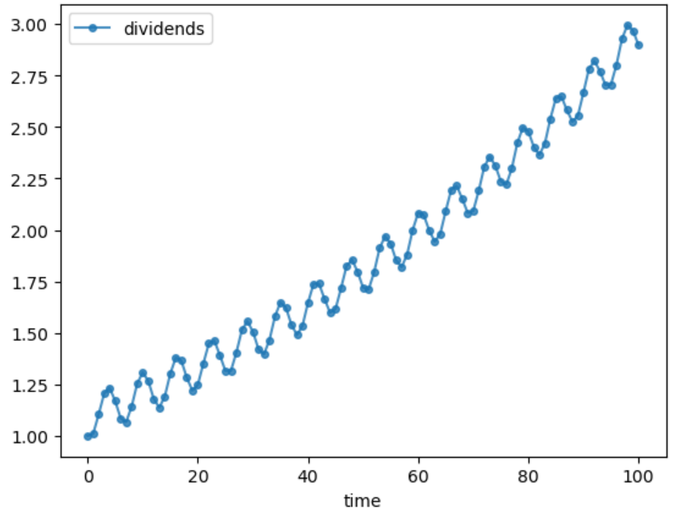
- Python 복습과 활용
- 기본적인 자료구조와 기초적 알고리즘 학습
- 실습
  - Gordon 기업성장모형
  - Gordon 주식가치 결정모형
  - 위험조정 성과지표 (Sharpe Ratio, Treynor Ratio, Jensen Ratio, Information Ratio)
  - 파생상품 차익거래
  - 이동평균선 도출

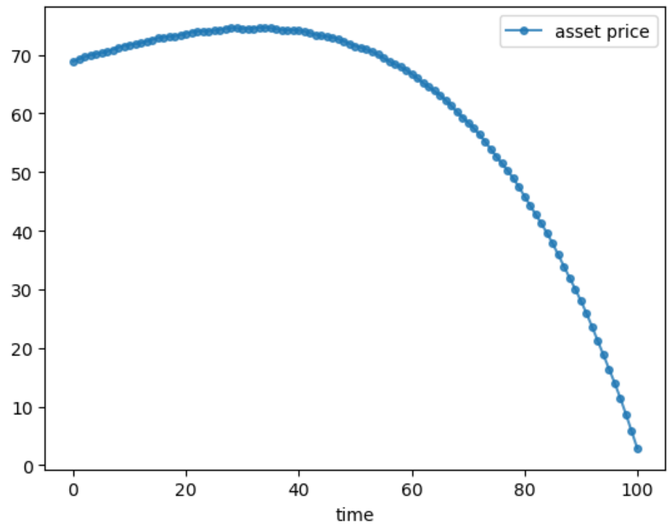


Numpy, Matplotlib, Linear Algebra

Week02

- Numpy
- Matplotlib
- Linear algebra for python
- 실습
  - Interest rate and present values
  - Discount factor
  - Gross Rate of Return

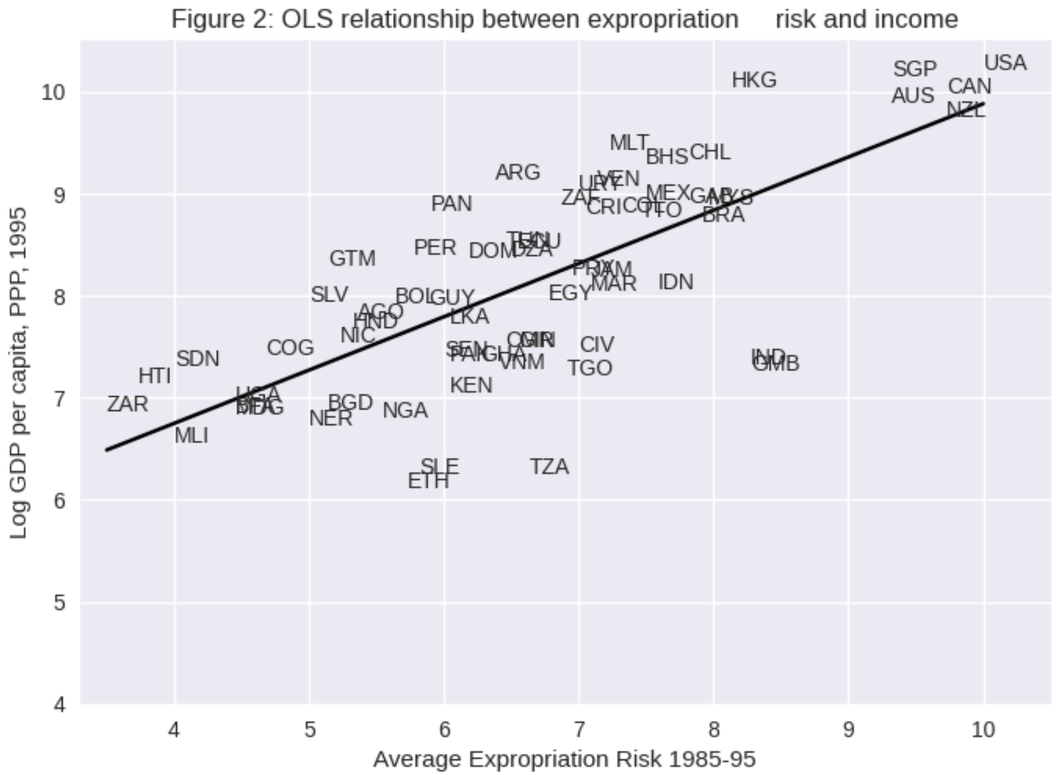




Pandas Visualization and Statistics

Week03

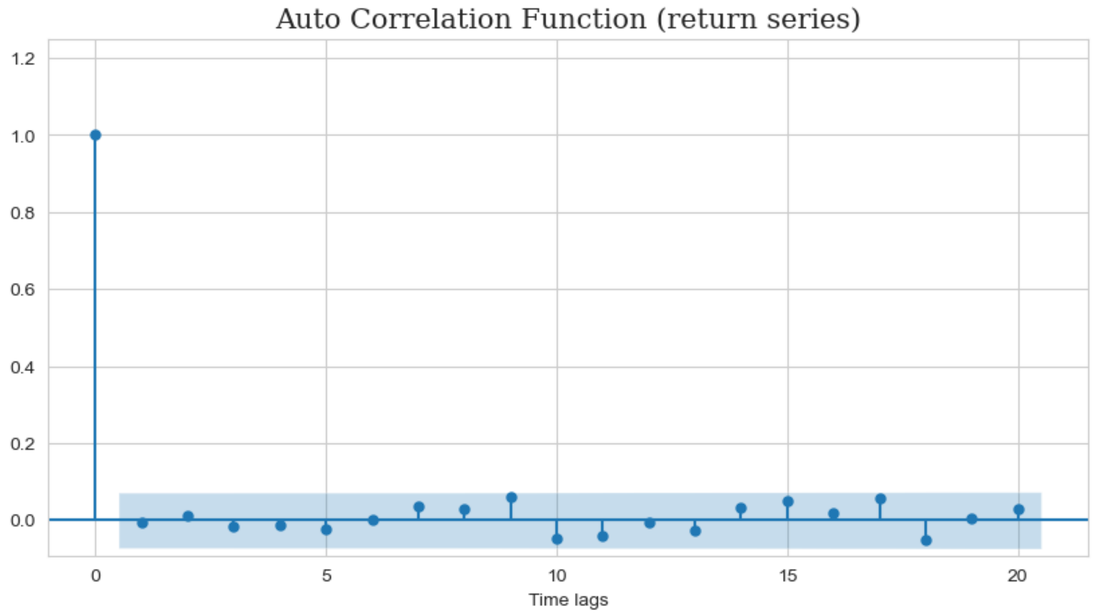
- Pandas
- matplotlib을 활용한 visualization
- 기초 통계학
- 실습
  - Simple Linear Regression
  - Multivariate Regression Models
  - Endogeneity



Stationary & Basic concepts of Time Series

Week04

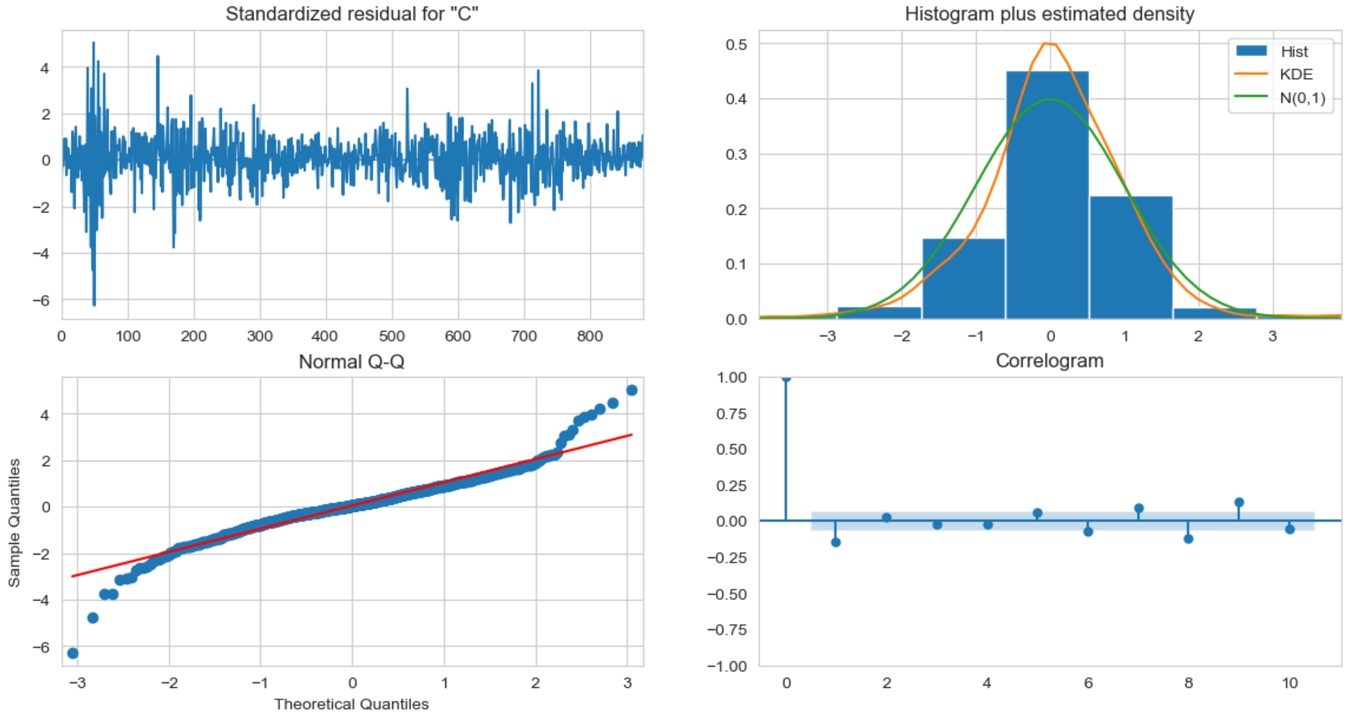
- Asset Returns
- Distributional Properties of Returns
- Stationarity
- 실습
  - 로그 차분 시계열
  - Autocorrelation
  - 자산간 상관관계



Univariate Time Series

Week05

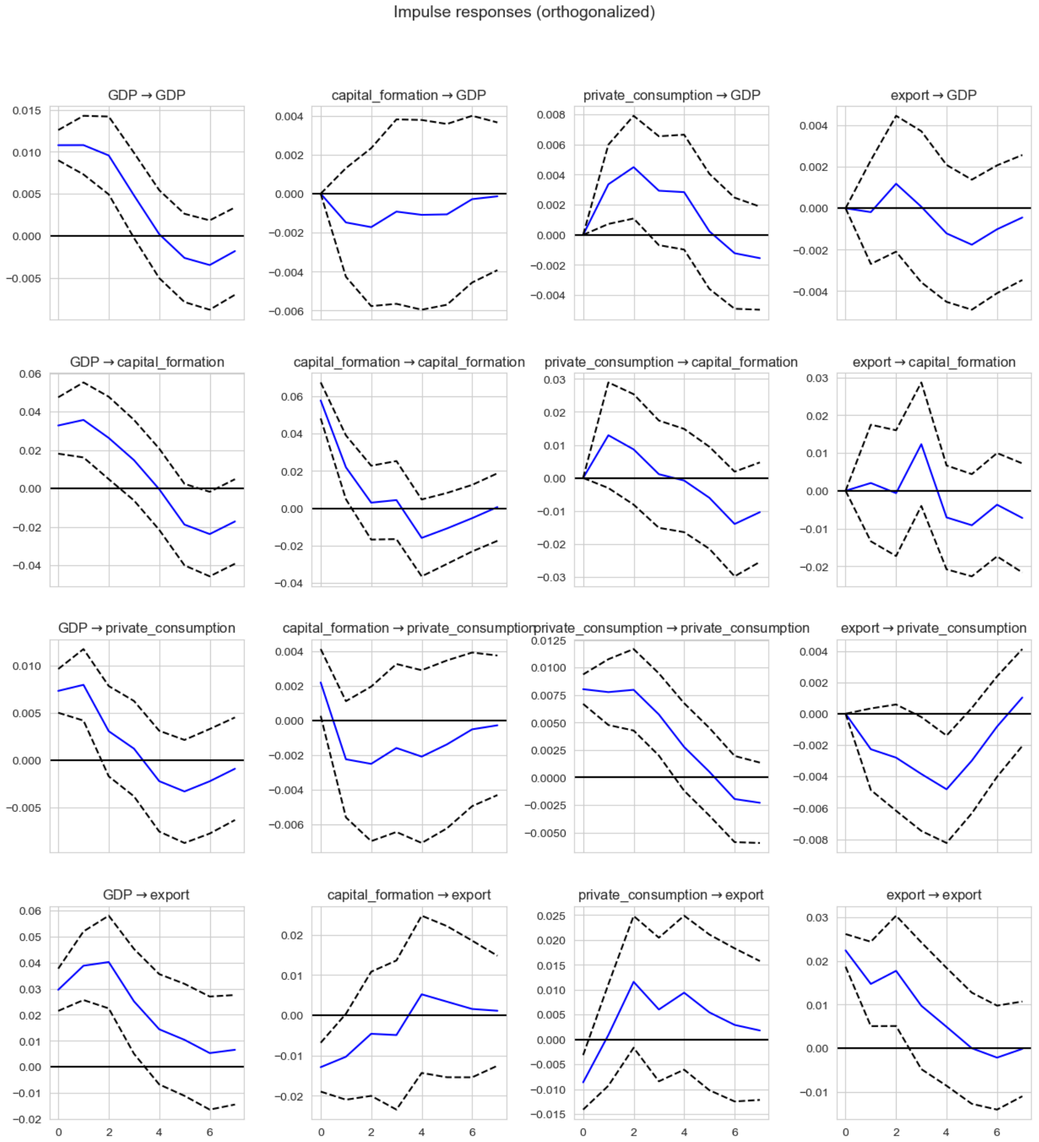
- Linear Time Series ; AR, MA, ARMA
- Linear Time Series Forecasting
- 실습
  - AR Modeling
  - MA Modeling
  - ARMA Modeling
  - Forecasting
  - Recurrent Neural Network



Multivariate Time Series

Week06

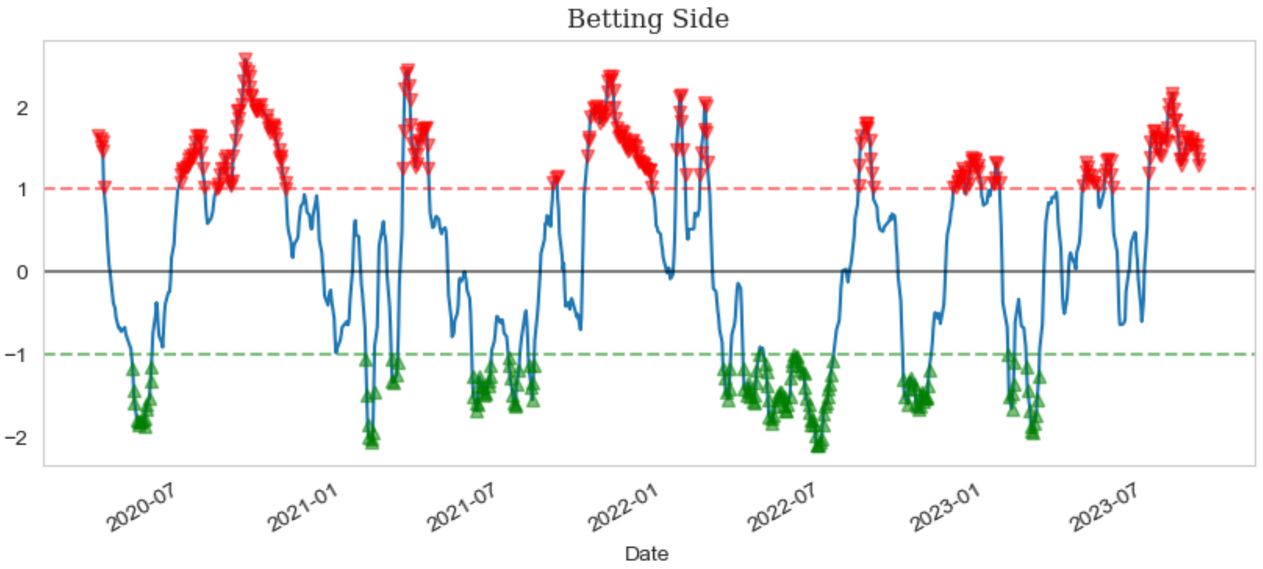
- Vector Process
- VAR
- VMA
- VARMA
- Impulse Response Function (IRF)
- Granger Causality
- 실습
  - CAPM
  - Beta Mapping risk estimation
  - GDP Affection
  - VAR Model, Kronecker Products

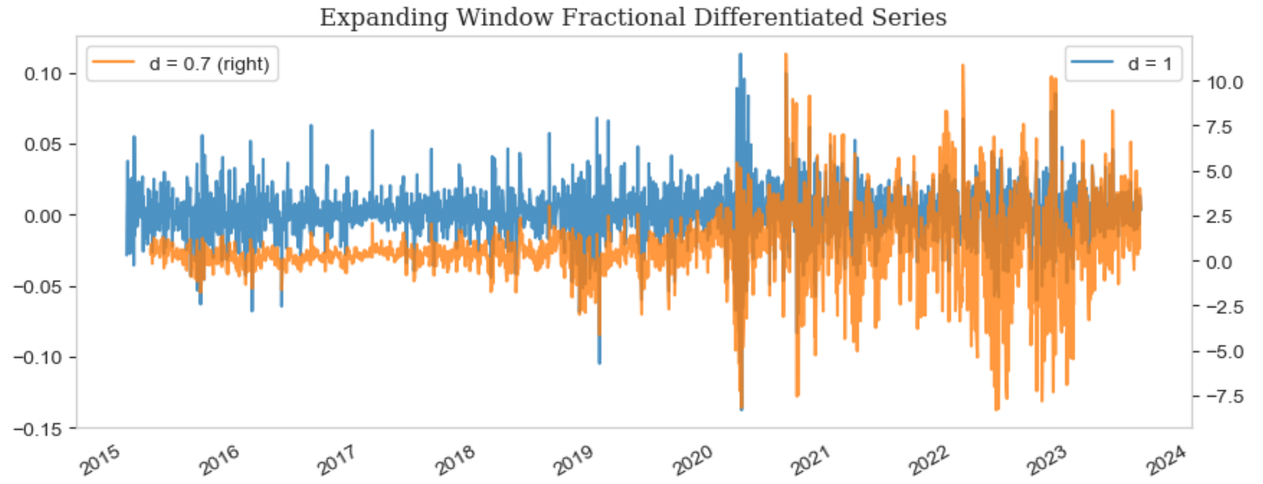


Integrated Time Series

Week07

- Integration of order
- Cointegration
- Trend Stationarity
- Long Memory
- 실습
  - Cointegration test and VAR Model
  - Pair Trading
  - Market Long memory

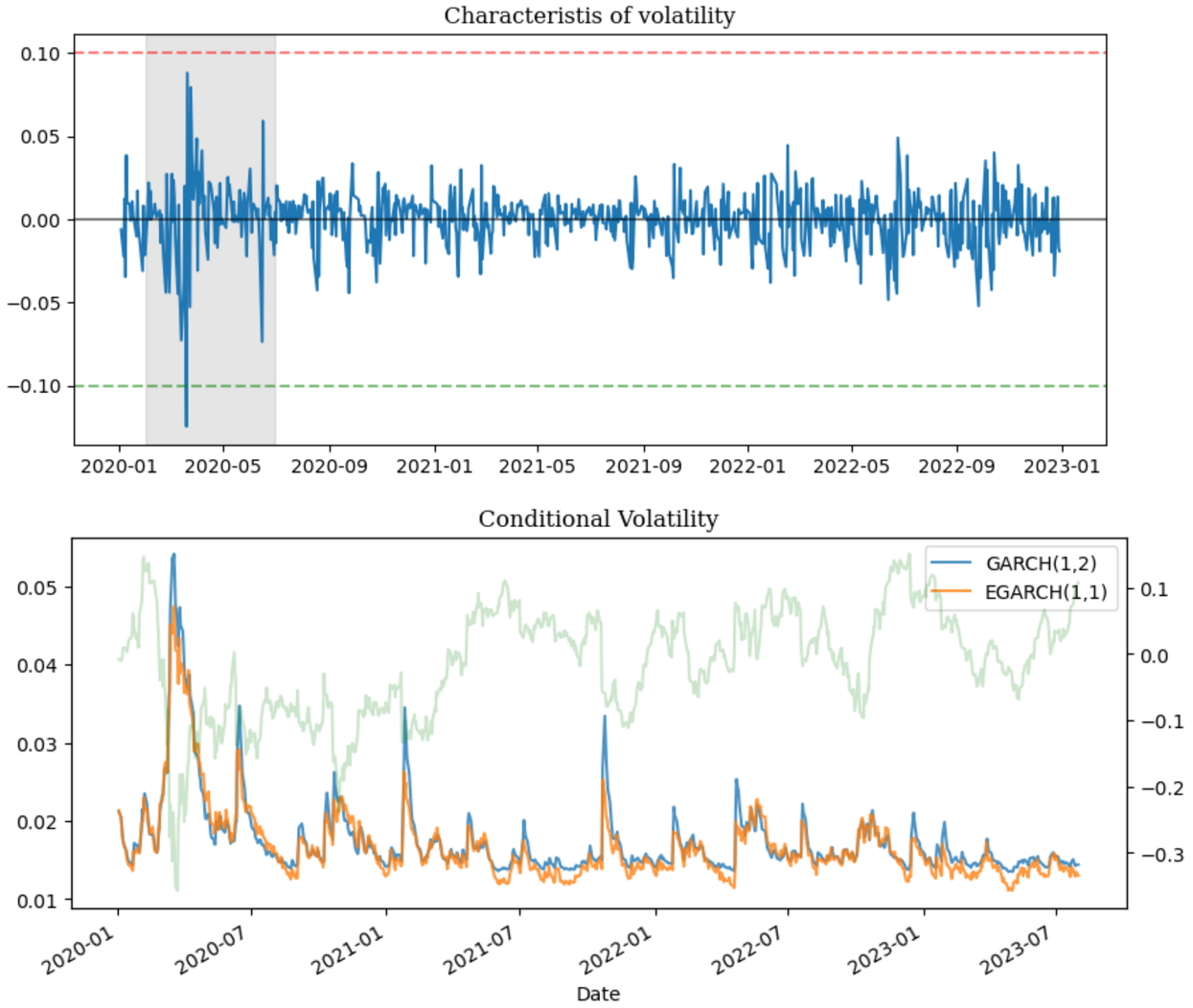




Multivariate Time Series

Week08

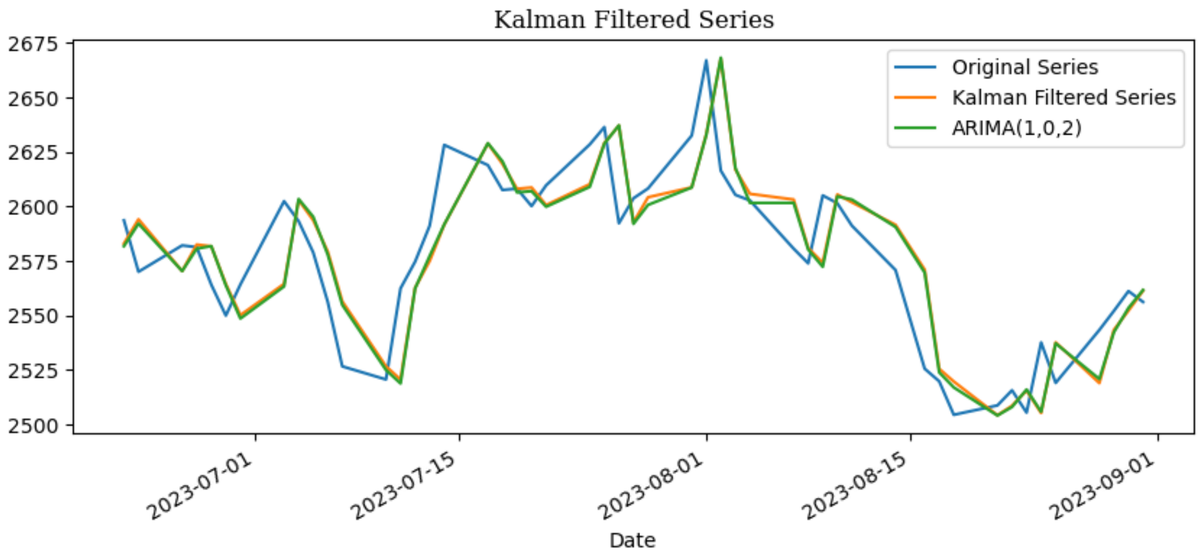
- Heteroscedasticity
- Volatility Clustering
- ARCH model
- GARCH model
- 실습
  - GARCH, EGARCH
  - Value at Risk(VaR)
  - EWMA Volatility

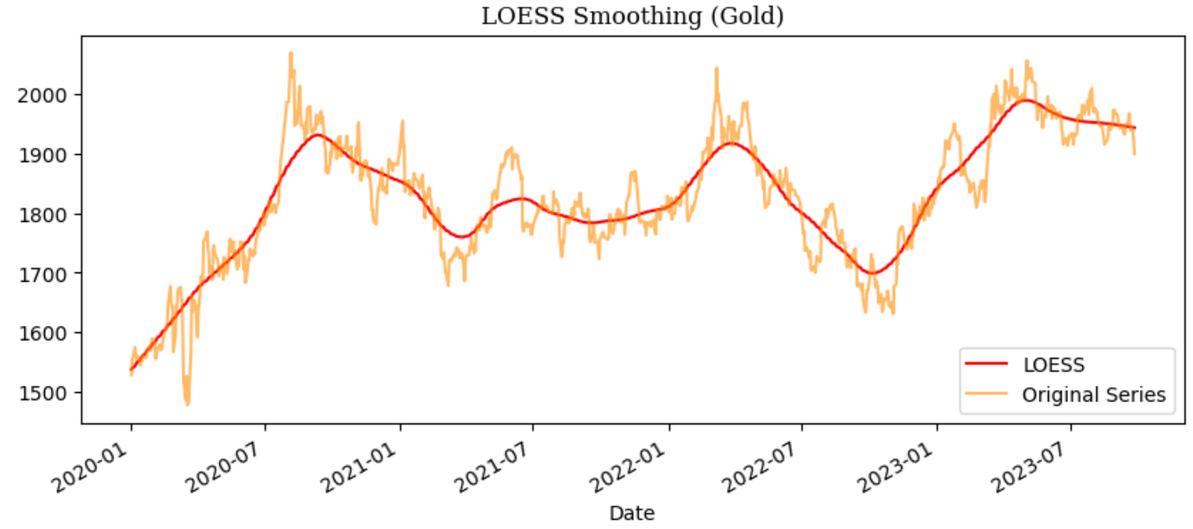


State Space Models

Week09

- Kalman Filter
- Kalman Smoother
- 실습
  - State Space Model
  - Kalman Smoothing
  - LOESS Smoothing (kernel smoothing)

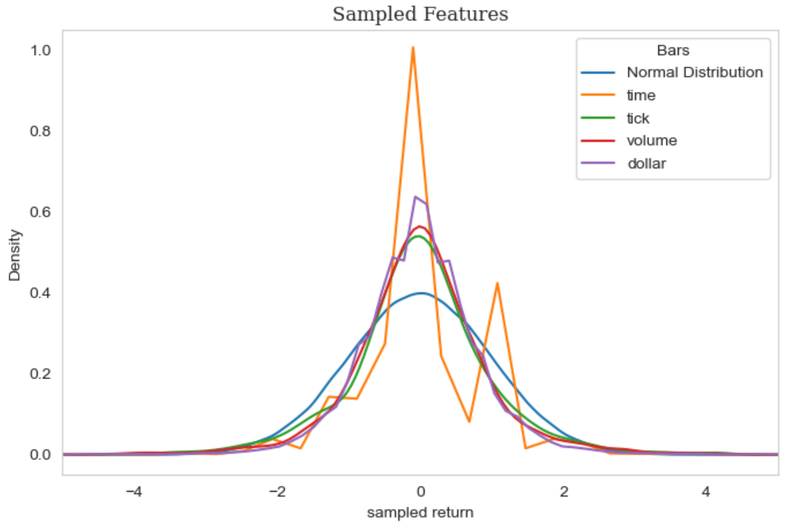


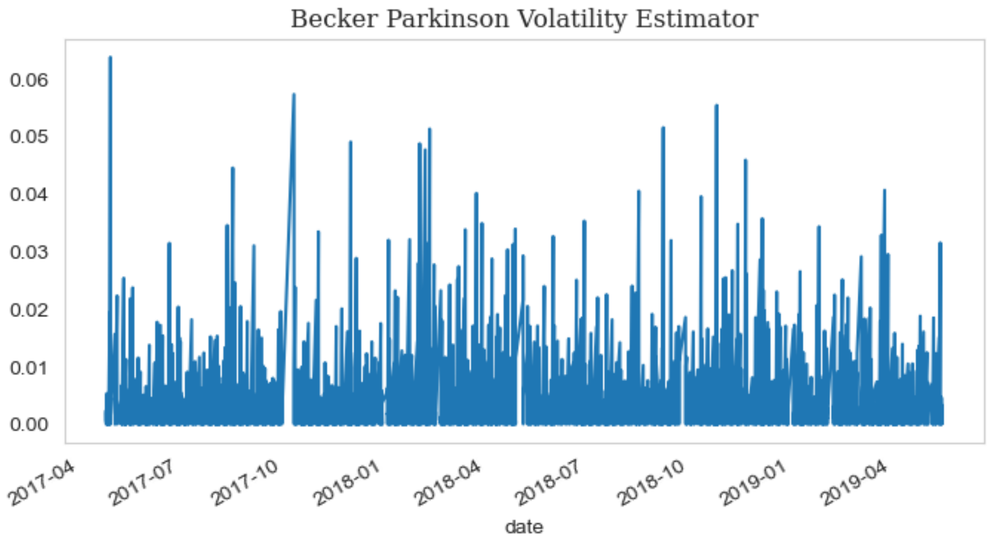


High Frequency Data Anala

Week10

- Price Sequences (Tick rule, Roll model, Corwin and Schultz)
- Volatility estimator ; Becker Parkinson
- Bar sampling
- 실습
  - Samsung Electronics 1min data analysis
  - Dollar value bar sampling
  - Roll modeling
  - Becker-parkinson volatility estimator

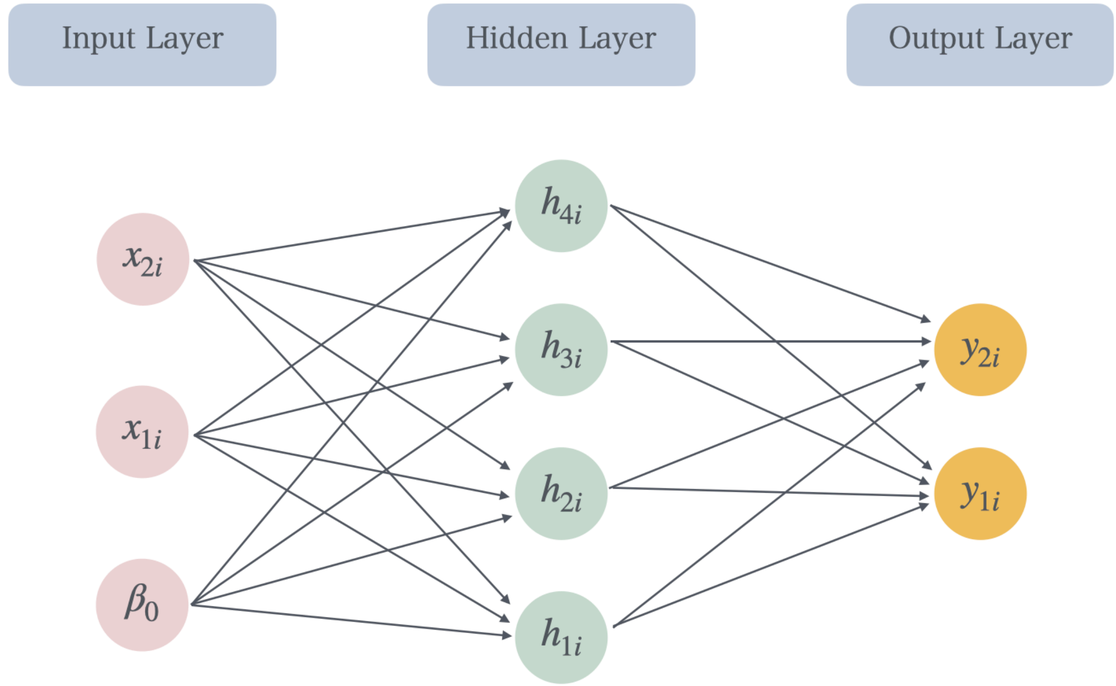




Concepts of Machine Learning

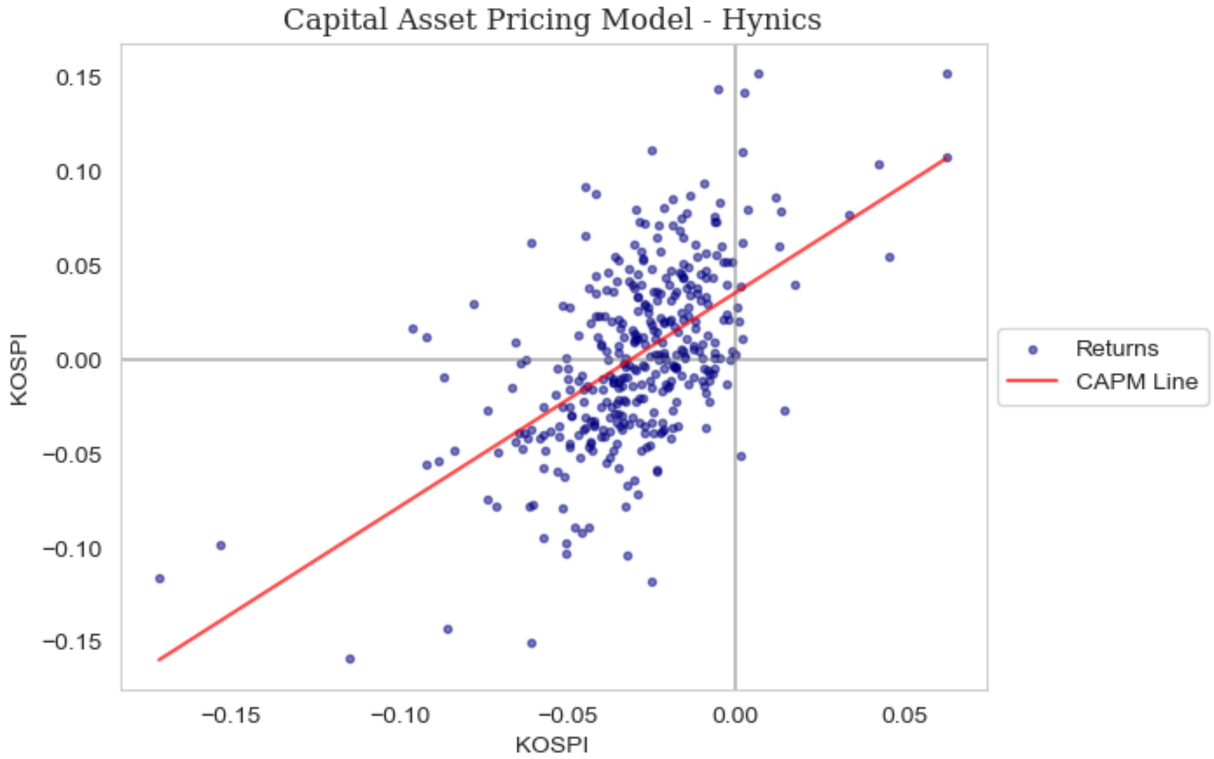
Week11

- OLS Models
- Multiple Linear Regression
- Classification and metrics
- Neural Networks
- Regularized Linear Models
- 실습
  - HAR Model
  - Ridge Regression
  - Lasso Regression
  - Elastic Net Regression



Regression
Week12

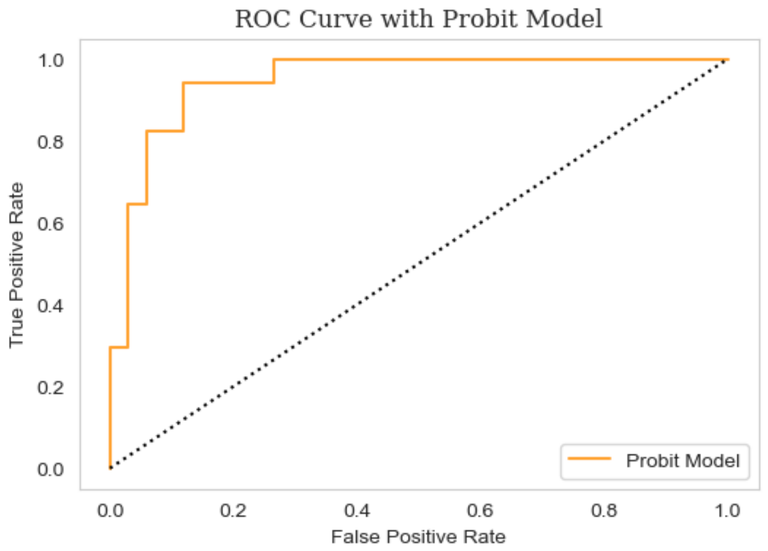
- Bias Variance Dilemma
- Support Vector Regression
- Regression Tree
- 실습
  - Security Market Line
  - Linear Regression ; Ridge
  - Support Vector Regression
  - Regression Tree

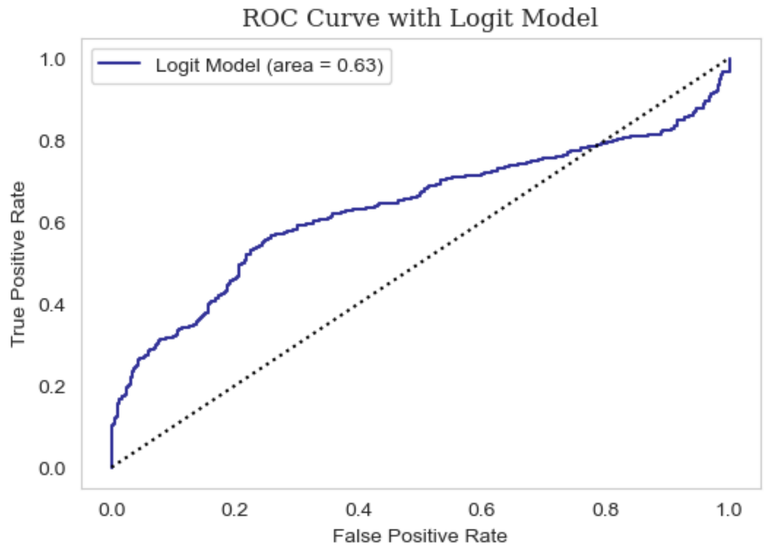


Classification

Week13

- Logit / Probit Models
- Naive Bayes
- Support Vector Machine
- Decision Tree
- 실습
  - Meta labeling
  - Market technical features
  - Primary model ; moving average strategy

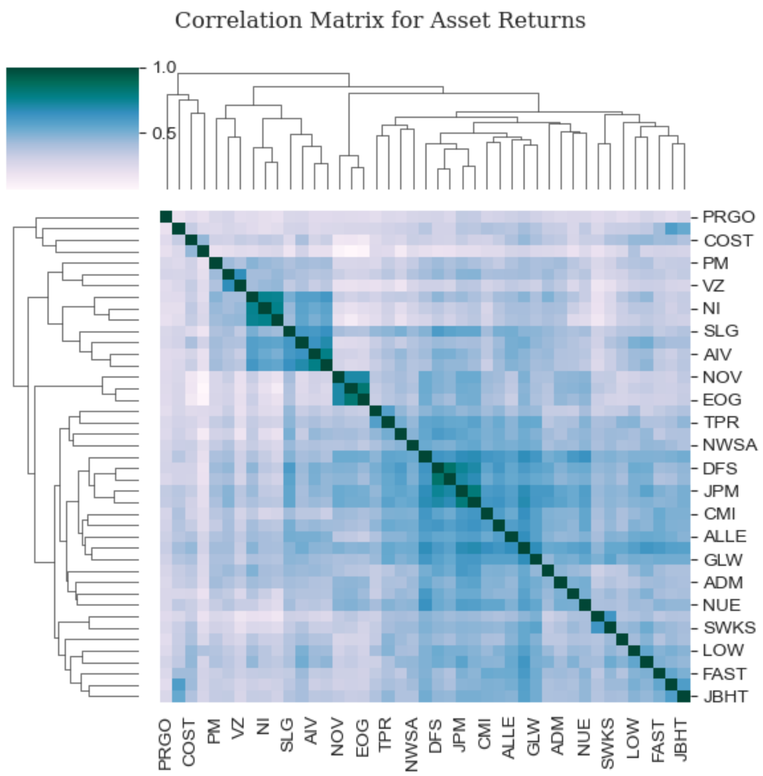


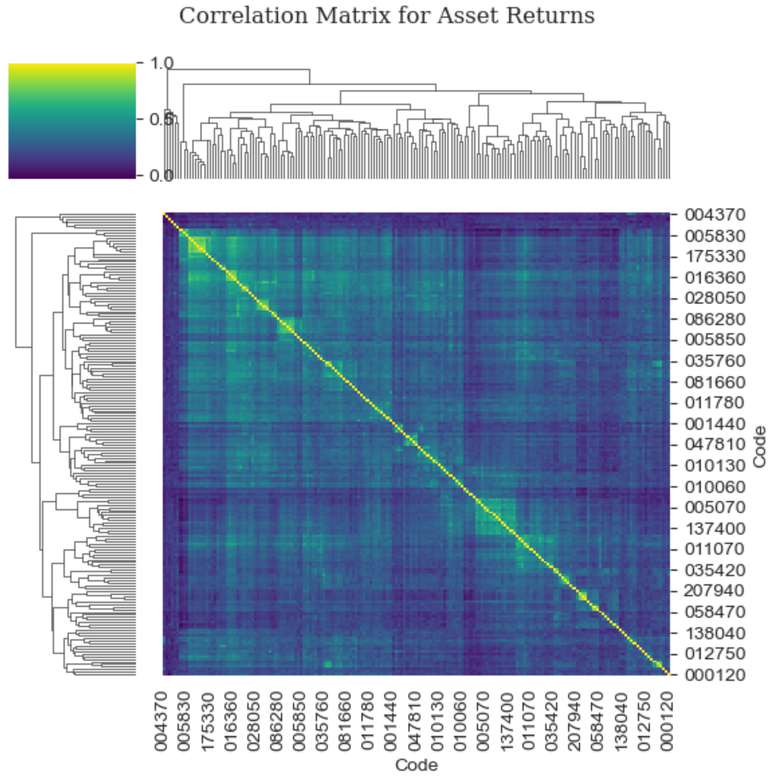


Dimension Reduction & Clustering

Week14

- Principal Components Analysis
- K-means Clustering
- DBSCAN Algorithm
- Hierarchical PCA allocation
- 실습
  - Asset Allocation
  - Quasi-diagonalization
  - Inverse PCA Allocation

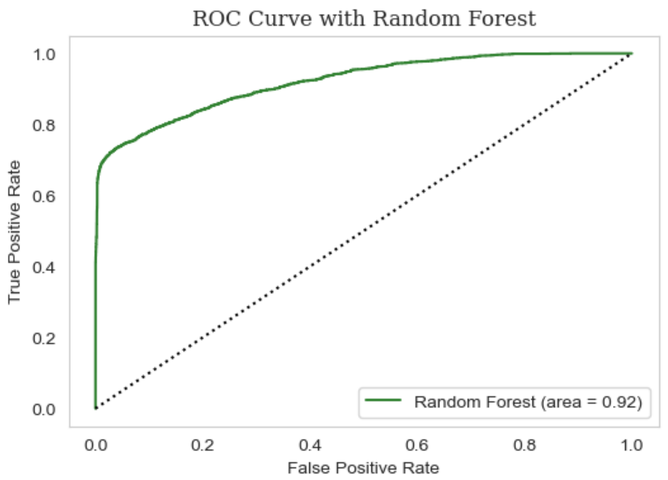


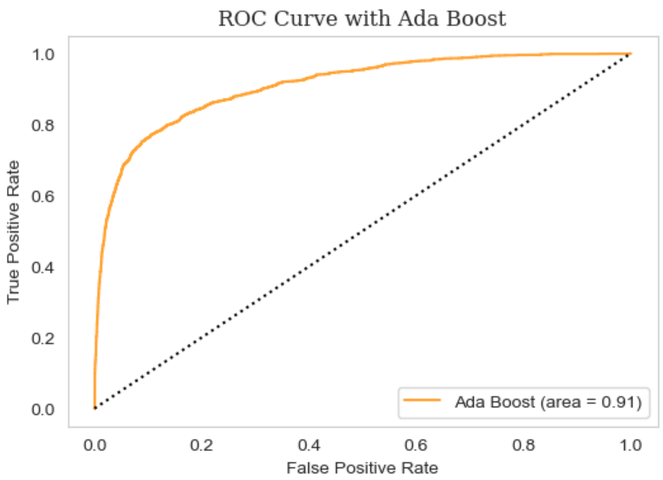


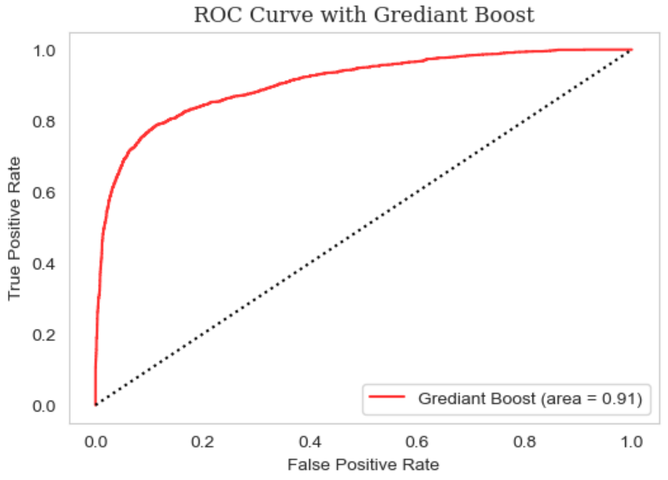
Ensemble Models

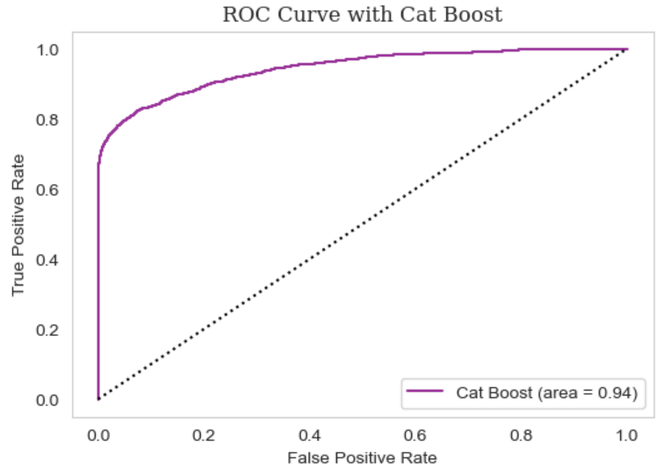
Week15

- Bagging, Boosting
- Random Forest
- Ada Boosting
- Gradient Boosting
- Cat Boosting
- 실습
  - Meta Labeling
  - Bagging
  - Boosting





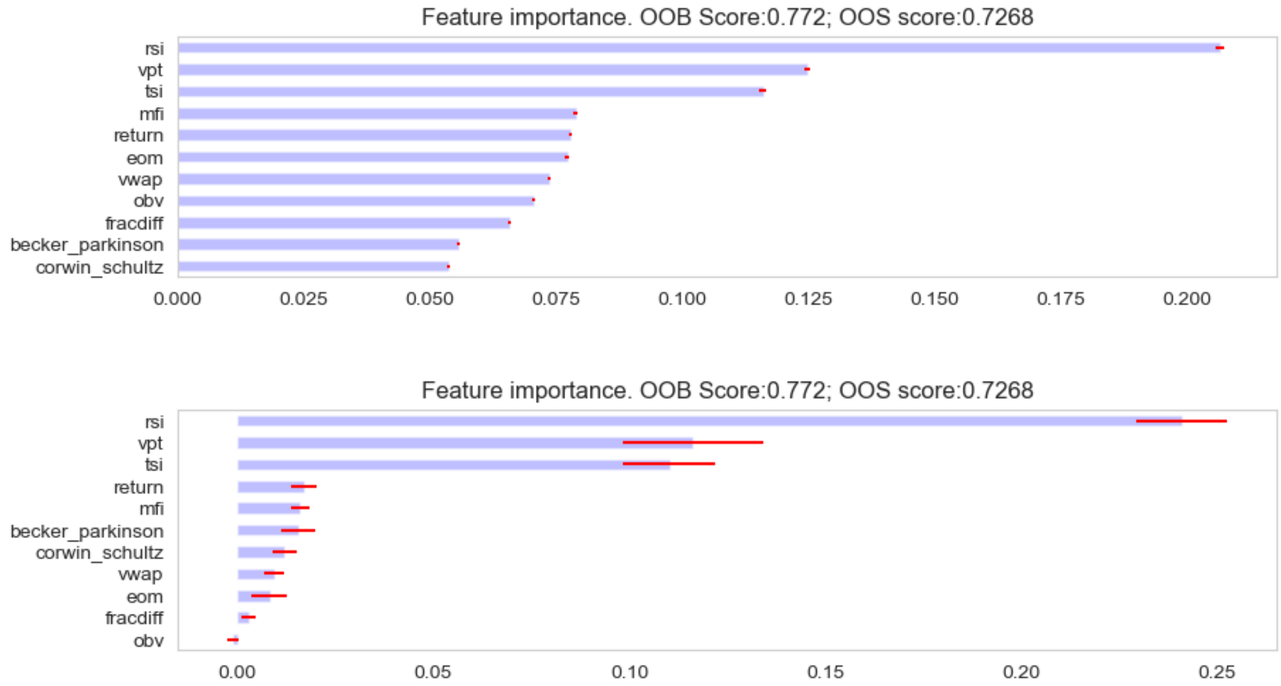




Explainable Machine Learning

Week16

- Feature Importances
- P-value
- MDA, MDI
- Probability weighted accuracy
- Shapley Value
- Causal Machine Learning
- 실습
  - Financial Machine Learning
  - Explainability of Feature importances



Recent Research

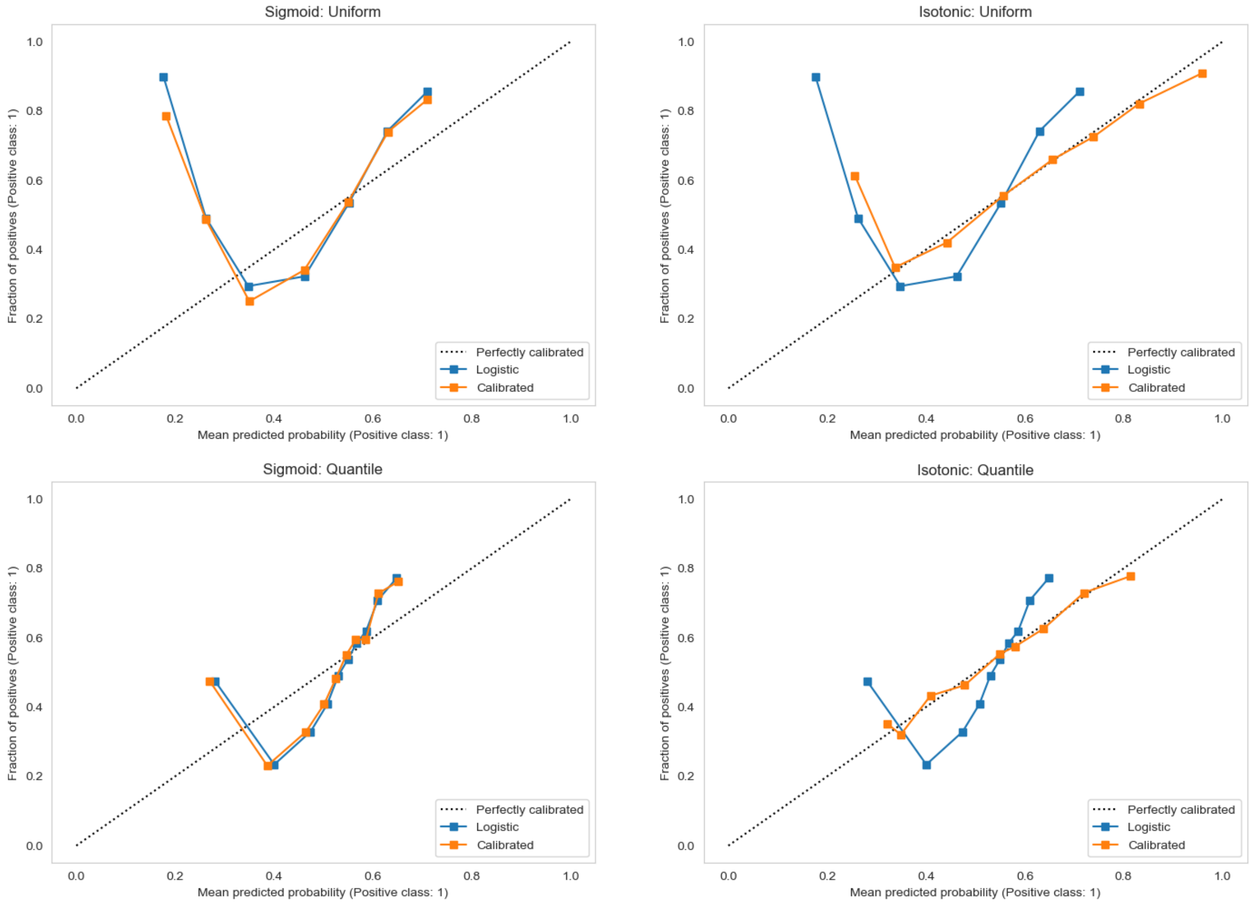
2024.01.17

Meta Labeling Strategies

Calibration and position sizing

Michael Meyer, Illya Barziy, and Jacques Francois Joubert (2023)

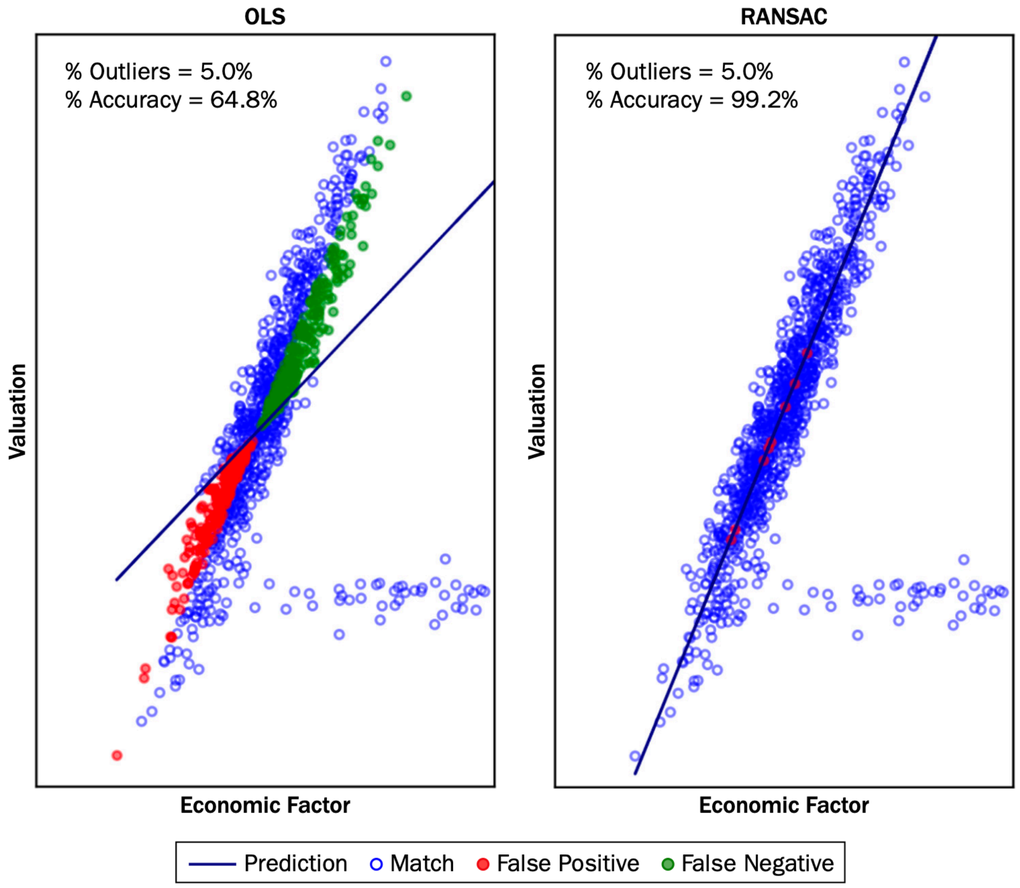
- Machine Learning의 probability는 편향된 확률
- 추정된 확률을 보정하고, 참값이 나오게 될 확률로 보정
- 보정한 확률을 기반으로 position size를 계산하면, 성과가 더 좋아질 것



Causation

Machine Learning for econometrics, Lopez de Prado(2022)

- ML offers the opportunity to gain insight from new datasets that cannot be modeled with econometric methods and old datasets that incorporate complex relation- ships that are still unexplored
- Existence : ML has been used to evaluate the plausibility of a theory across all scientific fields, even beyond the empirical sciences.
- Importance : ML algorithms can determine the relative informational content of variables
- Causation : ML algorithms can search for the structure of causal relation- ships, by analyzing statistical properties of purely observational data
- Reductionist : ML techniques are essential for the visualization of large, high-di- mensional, complex datasets
- Retriever : ML is used to scan through big data in search of patterns that humans failed to recognize



학회 활동


1조 팀프로젝트 진행상황 발표


Market Regime Change
Is it available for Machine Learning Features?
2023.09.06
Tommy Lee
ALBATROSS
SOGANG UNI.
American Football Team

학회 활동

Session Summary

회의록 + Filter Sort New

Aa 세션주차	답안제출	질문공유	박영우	엄홍재	한영태	이재희
Week 01	YES	YES	Prac01Unders...	엄홍재1.i...	Prac01U...	Prac01U...
Week 02	YES	YES	prac02_sol.pdf Prac02Present...		Prac02P...	
Week 03	YES	YES	Prac03LinearR...	Prac03Li...		Prac03Li...
Week 04	YES	YES	Prac04AssetR...	Prac04A...	Prac04A...	Prac04A...
Week 05	YES	YES				Prac05P...
Week 06	YES	YES	Note06Multiva...			Prac06V...
Week 07	YES	YES	Prac07PairTra...	Prac07P...	Prac07P...	Prac07P...
Week 08	YES	YES	Prac08GARCH...	Prac08G...		
Week 09	YES	YES		Prac09Fi...		
Week 10 OPEN	YES	YES				
Week 11	YES	YES				
Week 12	YES	YES				
Week 13	YES	YES				
Week 14	YES	YES				
Week 15	YES	YES				
Week 16	NO	YES				

+ New

Calculate

No	Aa 이름	Files & media	Number
03	iShare500 S&P500 future ticks	IVE_tickbidask...	HighFrequenyData Future ETF
04	KOSPI200 ticks	kospi_sample_...	HighFrequenyData Future
05	KOSDAQ150 ticks	kosdaq_sampl...	HighFrequenyData Future
06	Samsung Electonics	005930.txt	HighFrequenyData Stock
07	KOSPI index	kospi_1m.csv	HighFrequenyData Index
08	KRX Korean Stock List	stock_list.csv	Stock
09	monthly commodity prices	Commodity.xlsx	MonthlyData Commodity
10	Commodity-Gold	Gold_USD.csv	DailyData Commodity
11	KRW/USD	USD_KRW.csv	DailyData Currency
12	Korean Stock Trade Data (15min)	KOR15min.zip	HighFrequenyData Stock
14	코스피 전종목 시계열	kospi_all_199...	Stock DailyData
15	Western Digital(WSD) tick bid ask	WDC_tickbida...	HighFrequenyData Stock
16	AAPL Intraday Time Series	AAPL_1min_fir...	IntradayData Stock
17	AMAZON Intraday Time Series	AMZN_1min_f...	IntradayData Stock
18	Microsoft Intraday Time Series	MSFT_1min_fi...	IntradayData Stock
19	META Intraday Time Series	META_1min_fi...	IntradayData Stock
20	TESLA Intraday Time Series	TSLA_1min_fir...	IntradayData Stock
21	SPDR S&P 500	SPY_1min_firs...	IntradayData ETF
22	Invesco QQQ Trust	QQQ_1min_fir...	IntradayData ETF
23	S&P 500 VIX Short-Term Futures	VXX_1min_firs...	IntradayData ETF Future
24	SPDR Dow Jones Industrials Average	DIA_1min_first...	IntradayData ETF
25	iShares MSCI Emerging Markets Index	EEM_1min_fir...	IntradayData ETF
26	Apple Inc. Tick		HighFrequenyData Stock
27	IBM Tick		HighFrequenyData Stock
28	FOREX USD/JPY_1min		IntradayData Currency
29	S&P500 Tick	SP500ticks.csv	HighFrequenyData Future
30	FOREX EUR/USD Tick	FOREX EURUS...	HighFrequenyData Currency
31	KOSPI 전종목 1min data	KOSPI_1min.zip	IntradayData Stock

활동 및 수상내역 (2023-02)

- QRAFT Technologies research
- 현 DB금융경제공모전 3팀 출전
- NH투자증권 빅데이터 경진대회 우수상 수상
- 2023 서강 융합기술 경진대회 사회과학부문 대상, 장려상 수상

- 동계세션 python for finance : 23명 신청 (1월 16일 기준)


Intermission
2024.01.17

Contents

1부

- 2023-02 총정리
- 최근 Quantitative Finance 분야 연구 소개
- 학회 활동 및 수상내역 소개

2부

- 독립된 주제로써의 금융 머신러닝 (Financial Machine Learning) 연구를 위한 python library 제작
- Probabilistic Sharpe Ratio, Deflated Sharpe Ratio를 통한 Profit taking / Stopping loss threshold 지정
- Algorithmic Trading with Metha labeling strategies
- QA Chatbot

Developing
Financial Machine Learning
Library

2024.01.17

기존 연구에서의 문제점

금융에서의 머신러닝 프로젝트는 왜 실패하는가?

기존 연구에서의 문제점

파편화된 시스템

- 마구잡이로 선택되는 방법론
- 파편화된 python library
- 공학 연구에 목적을 맞춘 기존 ML 방법론

기존 연구에서의 문제점

적합하지 않은 class 사용

- Overfitting을 고려하지 않은 알고리즘 학습
- 적합하지 않은 Class 사용 (ex : cost function 등)
- 금융 연구만을 위한 Tool의 부재

Financial Machine Learning Library

금융 머신러닝 프로젝트를 위한 python library

ML in Finance

Financial Machine Learning Library

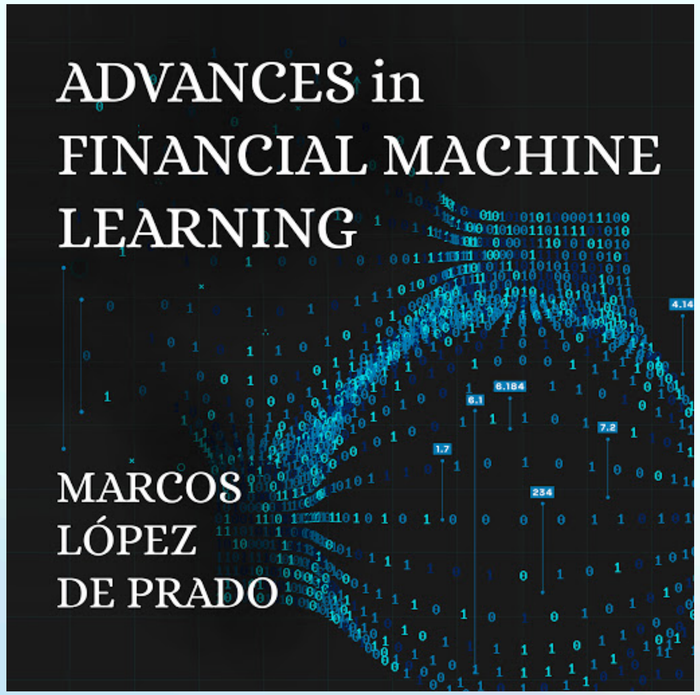
- 금융 머신러닝 연구를 위한 적절한 Tool 제공
- Overfitting을 고려한 Pipeline
- Feature Engineering을 위한 Sampling

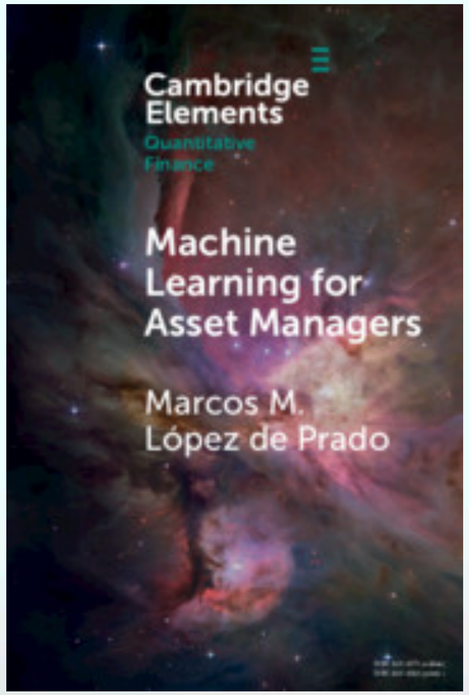
ML in Finance

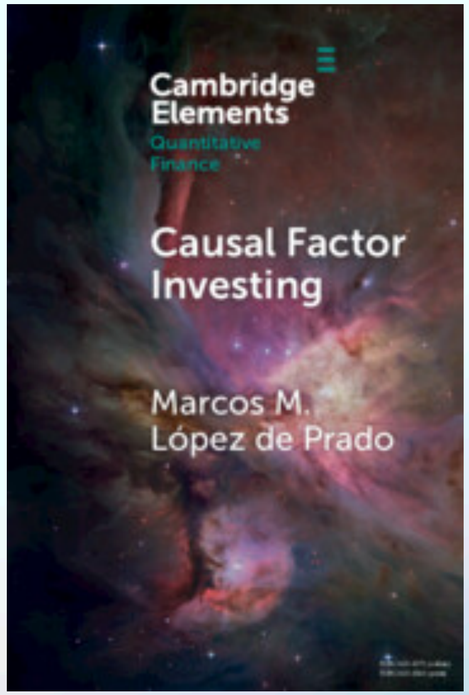
Financial Machine Learning Library

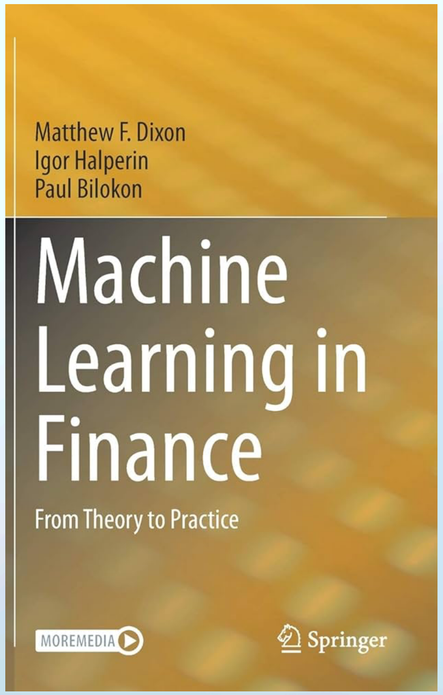
- ML/DL Algorithm과의 손쉬운 연동
- prediction 성능 향상을 위한 Smoothing
- Timeseries data 잡음 제거를 위한 Denoising, Detoning

FML library


ADVANCES in
FINANCIAL MACHINE
LEARNING
MARCOS
LÓPEZ
DE PRADO


Cambridge
Elements
Quantitative
Finance
Machine
Learning for
Asset Managers
Marcos M.
López de Prado


Cambridge
Elements
Quantitative
Finance
Causal Factor
Investing
Marcos M.
López de Prado


Matthew F. Dixon
Igor Halperin
Paul Bilokon
Machine
Learning in
Finance
From Theory to Practice
MOREMEDIA
Springer

FML library

Include Filtering, Statsmodels, Stochastic Simulation

- Conditional Heteroscedasticity Models
- Stochastic Process Simulation
- Kalman, LOESS, RTS, CNN smoothing

FML library

Include Behavioral Features and Labeling

- Market behavioral features(OBV, RSI, Moving average… etc)
- Triple Barrier Methods and Meta Labeling
- Trend Searching Labeling

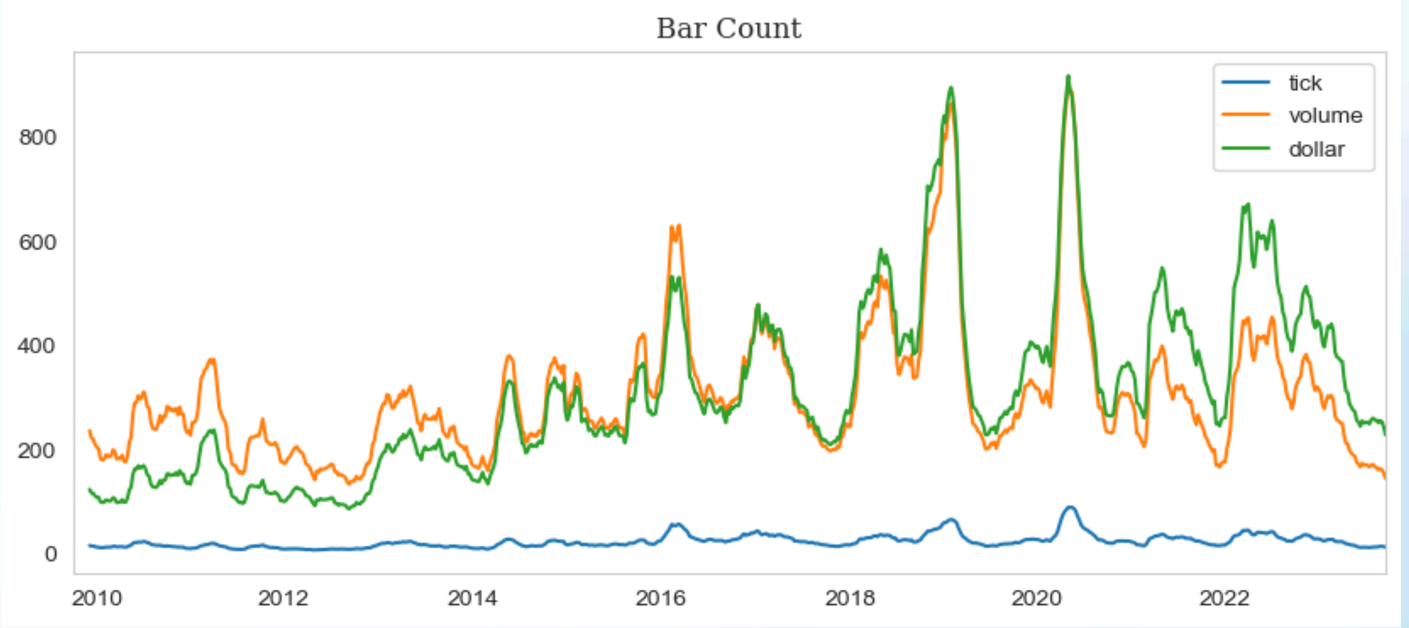
Financial Machine Learning Library

주요 기능 소개

FML library

Feature Engineering

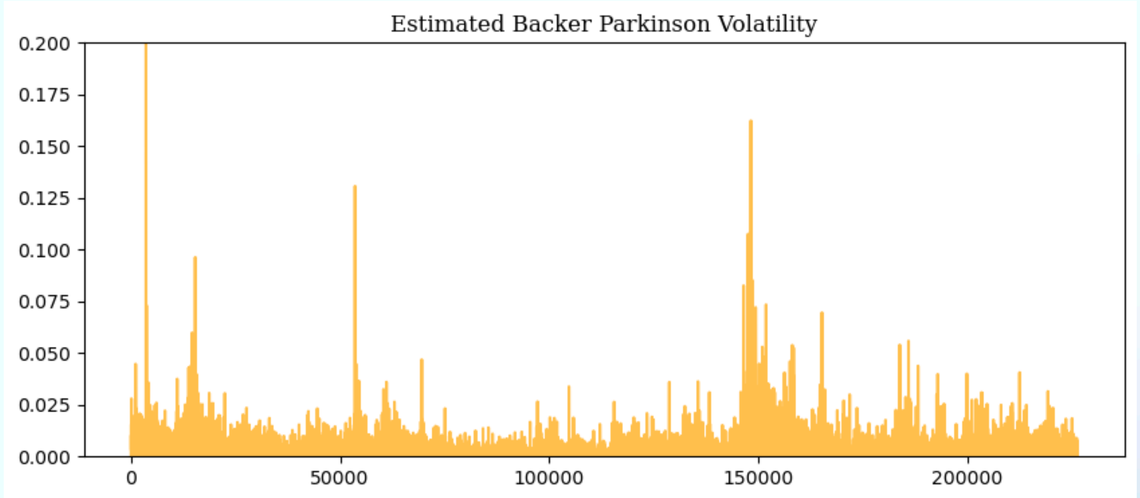
- High Frequency data feature engineering
- Standard bar, Imbalance bar, Run bar
- Can use ETF trick on futures series

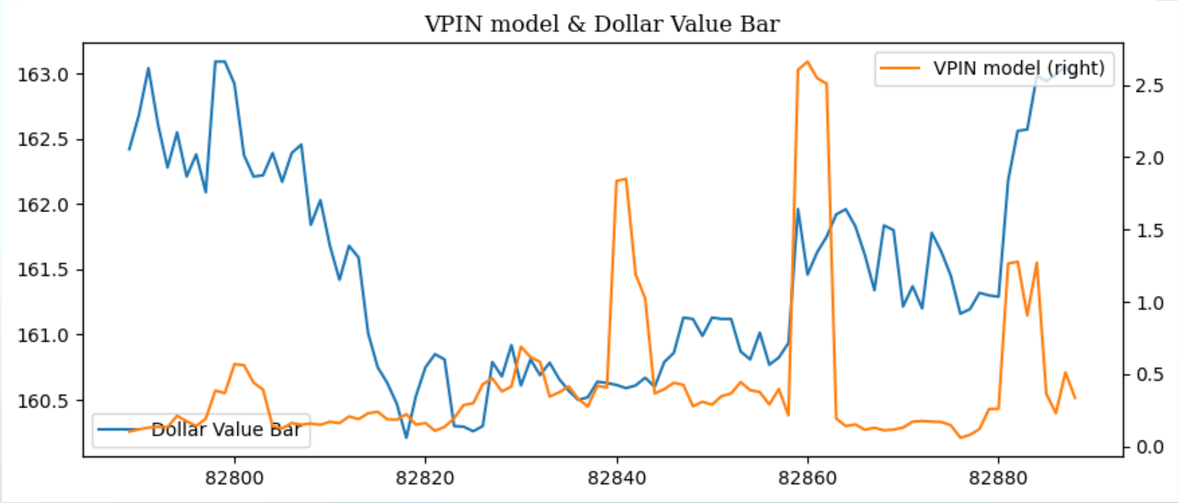


FML library

Market Microstructure Modeling

- Market Microstructural features
- Roll’s model
- Corwin-Schultz Spread
- Becker-Parkinson’s model
- VPIN
- Liquidity Lambda

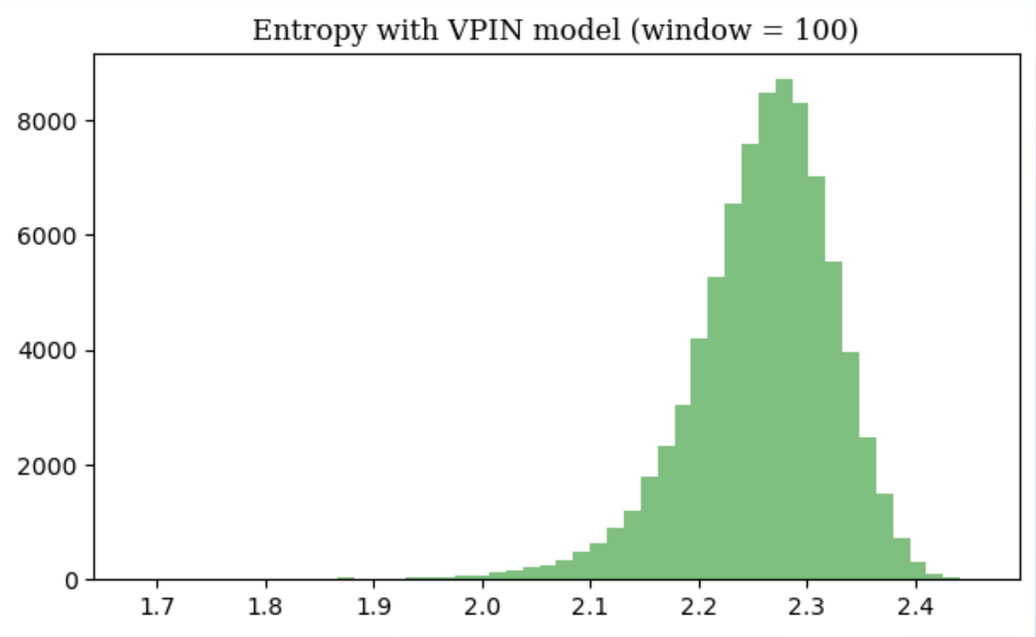




FML library

Useful features

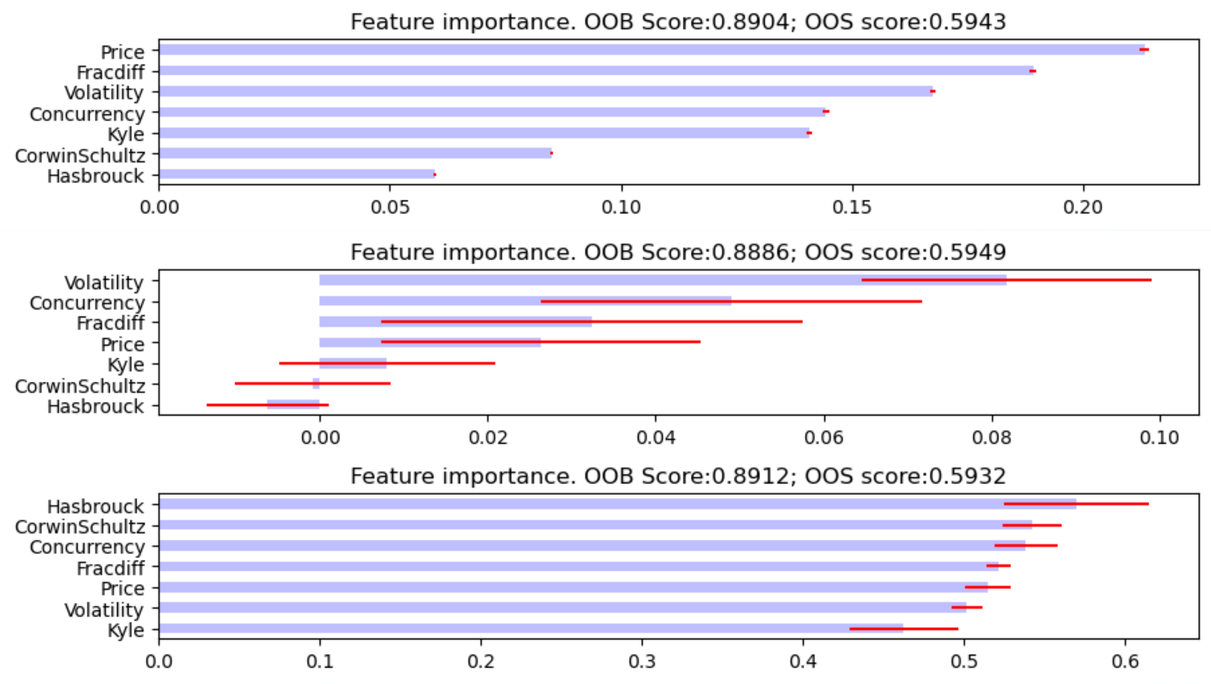
- Label Concurrency
- Fractionally differentiated features
- Entropy Features and estimator



FML library

Causal Inference Models

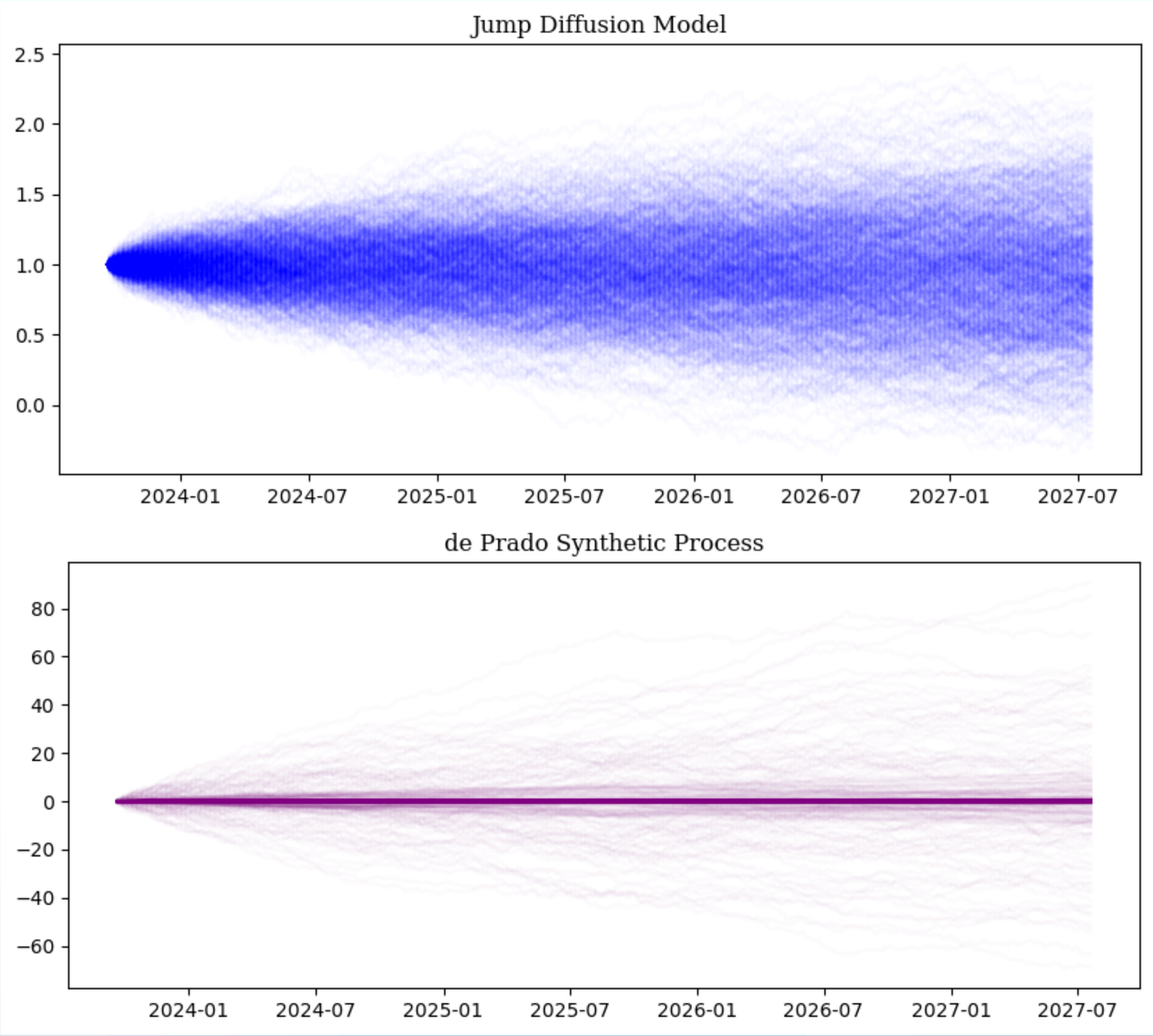
- All of Feature Importance is available
- Ensemble Model Learning
- MDA, SFI, MDI



FML library

Stochastic Simulation

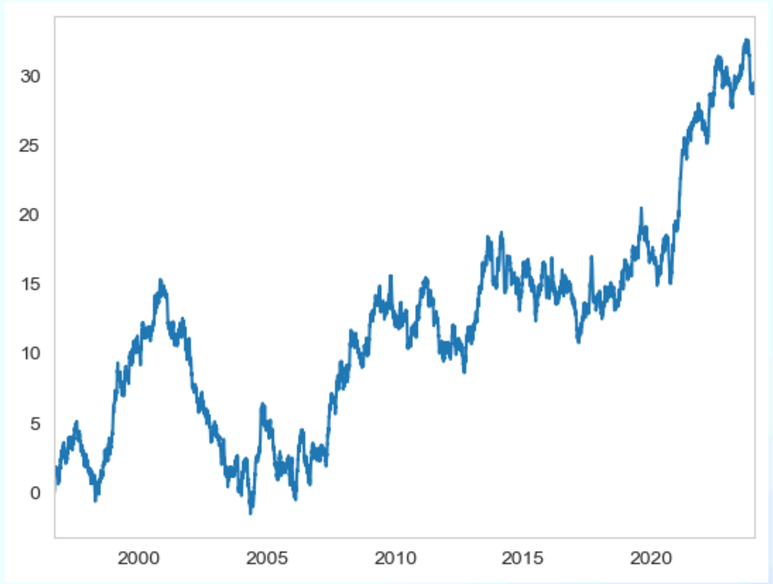
- Geometric Brownian Motion
- Ornstein Uhlenbeck Process
- Jump diffusion Models
- de Prado synthetic simulation
- AutoRegressive Modeling

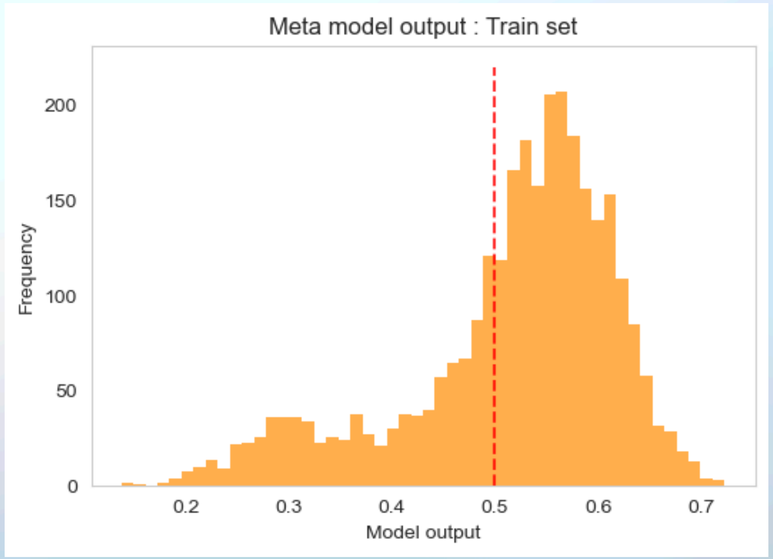


FML library

Stochastic Simulation

- Dual Regime
- Probabilistic Sharpe Ratio
- Betting sizing algorithms

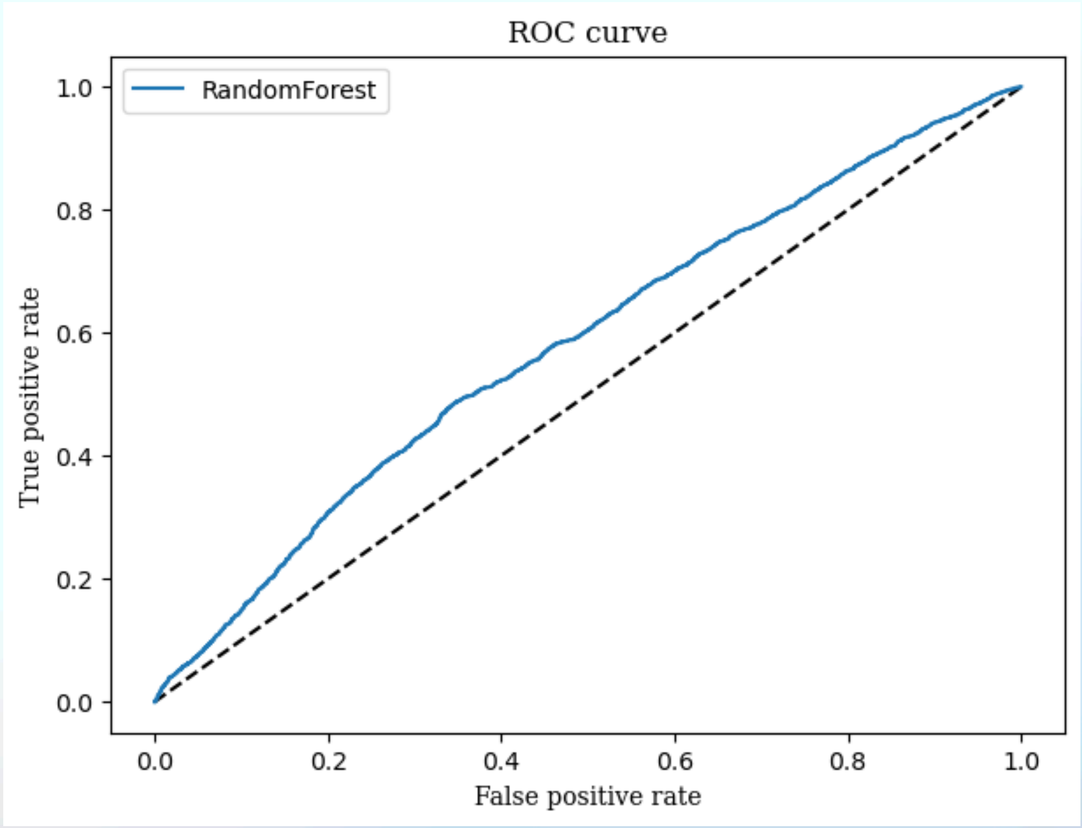




FML library

Meta Labeling Framework

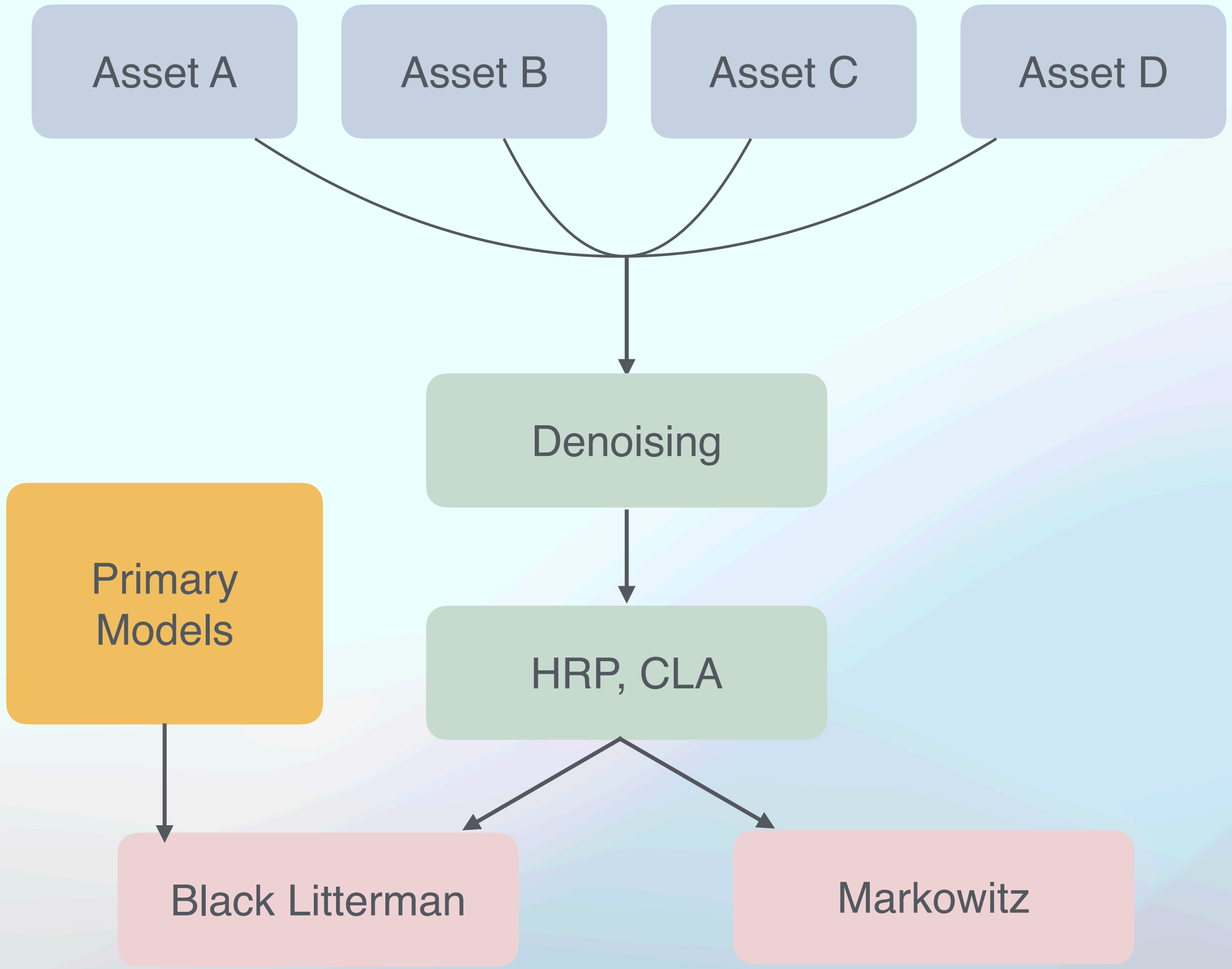
- Primary Model
  - Triple Barrier Method
  - EF3M method
  - Trand Search method
- Secondary Modeling
  - Hyperparameter Tuning
  - Bet Sizing, Prob Calibration



FML library

Machine Learning
Asset Allocation Pipeline

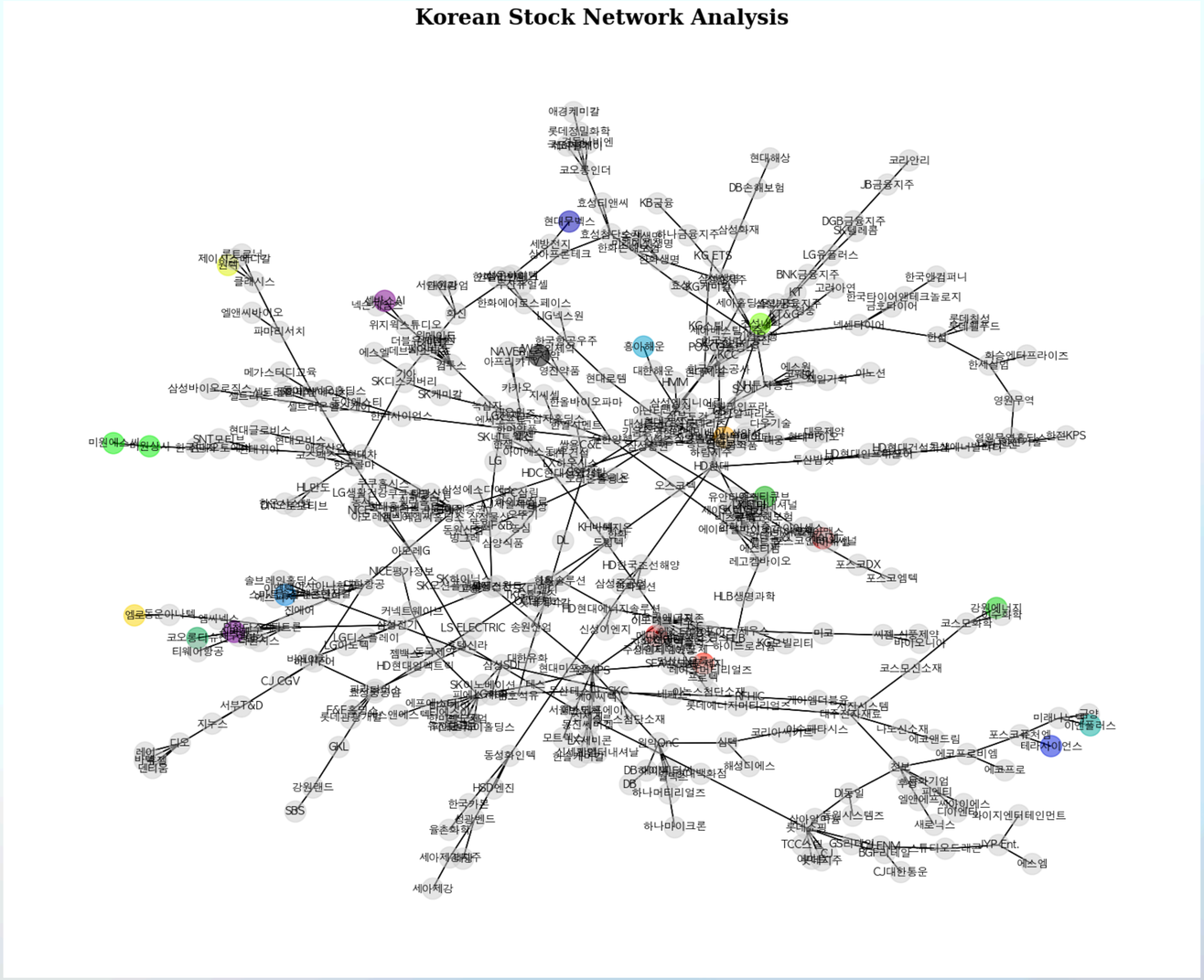
- Signal Denoising
- HRP, CLA, HCA optimization
- Black Litterman Optimization
- Powerful asset allocation tools



FML library

Deep Learning Concept Models

- Network Models
- CNN, Kernel Smoothing
- Autoencoder Simulation



왜 제작했는가?

ML 연구의 보편화

- 기존 계량경제학 모델링을 대신할 새로운 framework 필요성 대두
- 진입장벽을 낮추는 것에 목표
- 양질의 학회 연구를 위한 의도

왜 제작했는가?

편리한 사용성

- 이곳 저곳에서 끌어와 사용해야 하는 번거로움 해소
- Library 형태로 편리한 사용성 제공
- 연구에 있어서 일관된 결과 제공

Estimate PT / SL threshold using PSR and DSR

2024.01.17

Motivation

Sharpe Ratio 해석의 문제

- Walking forward 방식인 기존의 Sharpe ratio는 단지 수많은 시간 경로 중 하나의 경로만을 가지고 측정한 성과 기준
- 필연적으로 과적합의 문제를 가지고 있음
- 추정 parameter의 불안정성 문제

$$\text{Sharpe Ratio} = \frac{\hat{\mu}}{\hat{\sigma}}$$

- 문제점 : parameter는 실제로 모분산과 모평균에 근접하는가?

Probability Sharpe Ratio

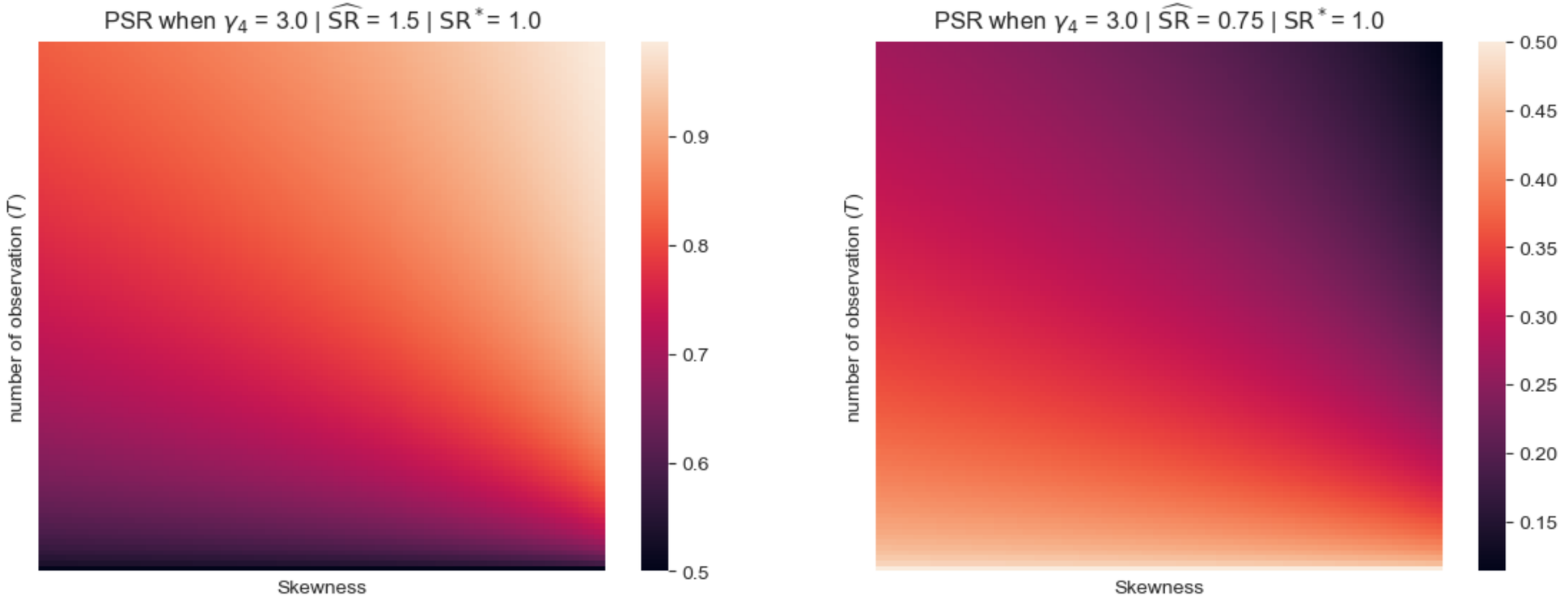
Monte Carlo simulation을 통한 확률적 SR의 추정

$$\widehat{\mathrm{PSR}}\left[\mathrm{SR}^*\right] = Z\left[\frac{(\widehat{\mathrm{SR}} - \mathrm{SR}^*)\sqrt{T-1}}{\sqrt{1 - \hat{\gamma}_3\widehat{\mathrm{SR}} + \frac{\hat{\gamma}_4 - 1}{4}\widehat{\mathrm{SR}}^2}}\right]$$

- $\widehat{\mathrm{SR}}$ 은 경험적 데이터로부터 추정된 표본 Sharpe Ratio
- SR^*은 목표 Sharpe Ratio
- $\hat{\gamma}_3$ 은 추정된 왜도, $\hat{\gamma}_4$ 은 추정된 첨도
- 확률적 Sharpe ratio를 위와 같이 추정이 가능. 즉, 하나의 sample로써 PSR을 추정

Probability Sharpe Ratio

Monte Carlo simulation을 통한 확률적 SR의 추정



- 왜도와 관측치가 증가할 수록 PSR이 증가하며, 첨도가 커질수록 PSR이 낮아지는 경향이 존재
- 5차 적률과 6차 적률을 포함한다면 어떻게 될까?

Backtest with Synthetic data

Monte Carlo simulation을 통한 Sharpe Ratio의 추정

- suppose a discrete Ornstein-Uhlenbeck (O-U) process on prices

$$P_{i,t} = (1-\varphi)\,\mathrm{E}_0[P_{i,T_i}] + \varphi P_{i,t-1} + \sigma\varepsilon_{i,t}$$

- the performance of opportunity i is characterized by the process

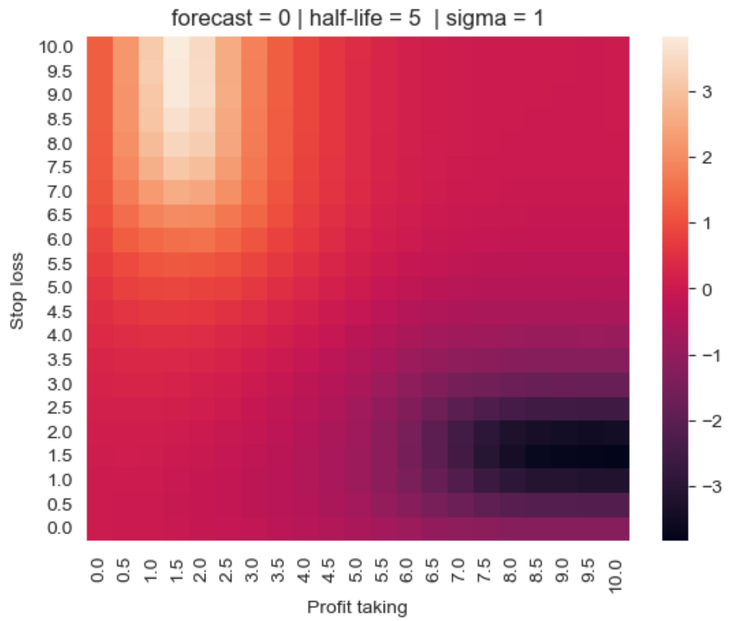
$$\frac{1}{m_i}\pi_{i,t} = (1-\varphi)\mathrm{E}_0[P_{i,T_i}] - P_{i,0} + \varphi P_{i,t-1} + \sigma\varepsilon_{i,t}$$

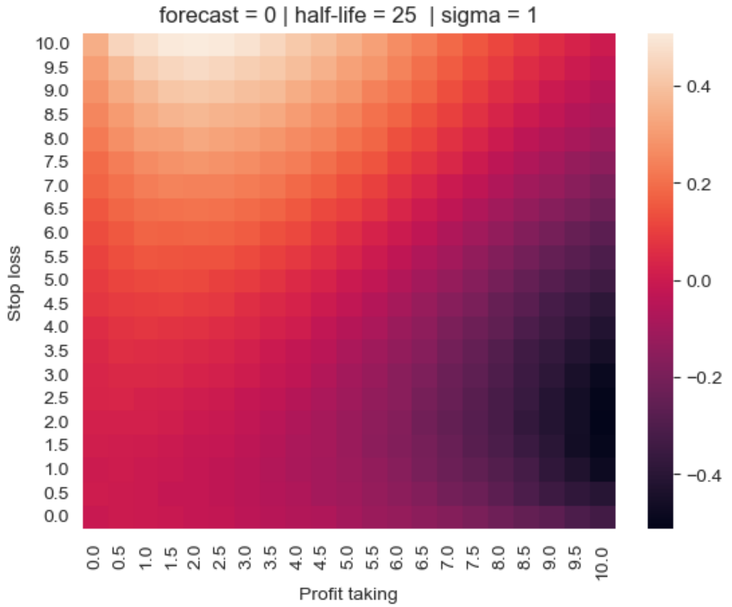
$$\pi_{i,t} \sim N\left[m_i\left((1-\varphi)\,\mathrm{E}_0[P_{i,T_i}]\sum_{j=0}^{t-1}\varphi^j - P_{i,0}\right), m_i^2\sigma^2\sum_{j=0}^{t-1}\varphi^{2j}\right]$$

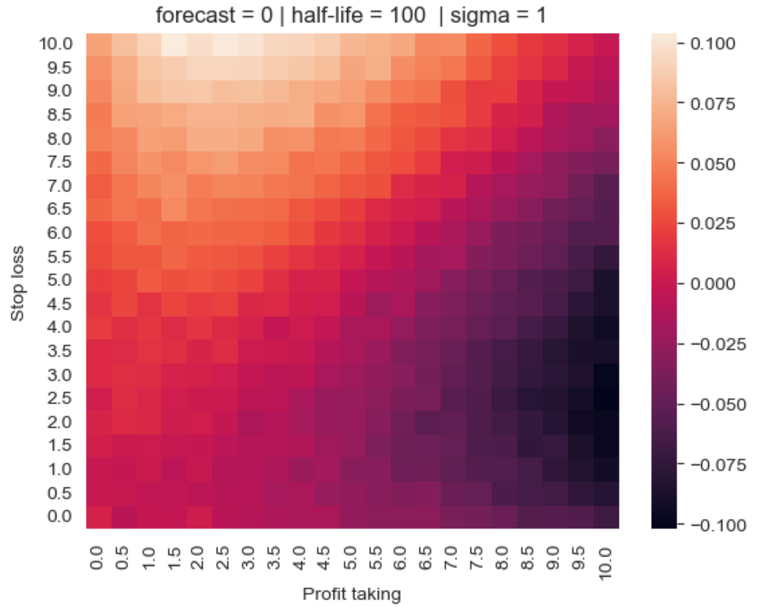
- The half life is $\tau = -\dfrac{\log(2)}{\log(\varphi)}$

Backtest with Synthetic data

Monte Carlo simulation을 통한 Sharpe Ratio의 추정



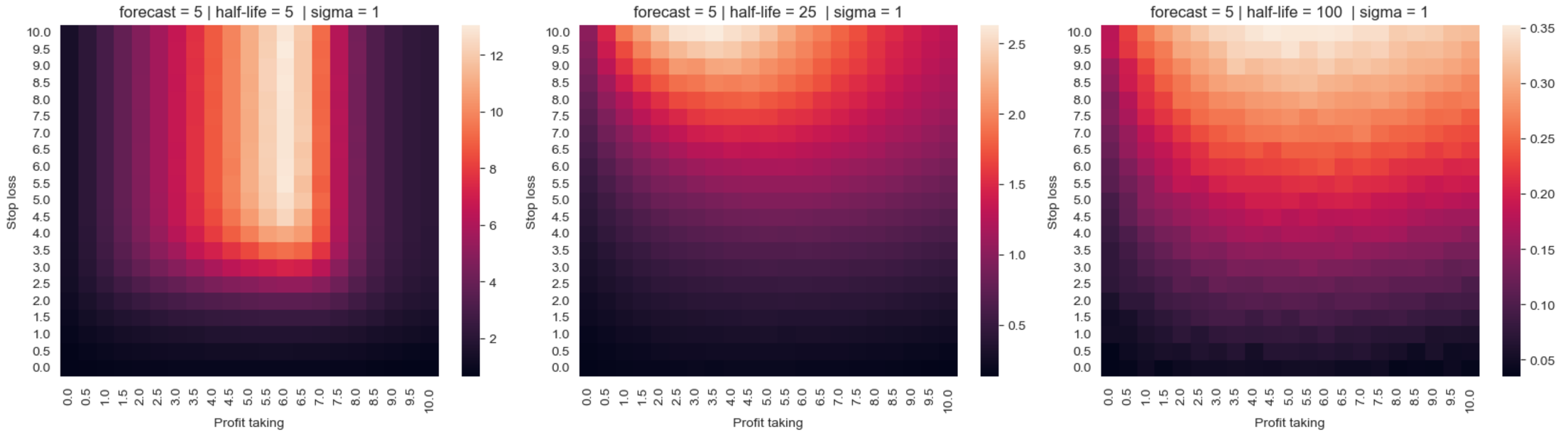




- 장기 균형이 0 즉, 횡보세를 보이는 시장의 경우, profit taking의 폭이 적고, stopping loss의 폭이 큰 지점에서 sharpe ratio가 극대화 됨
- 횡보장에서는 빠른 이익 실현과 손실시 보유하는 전략이 profit taking을 할 가능성이 높음

Backtest with Synthetic data

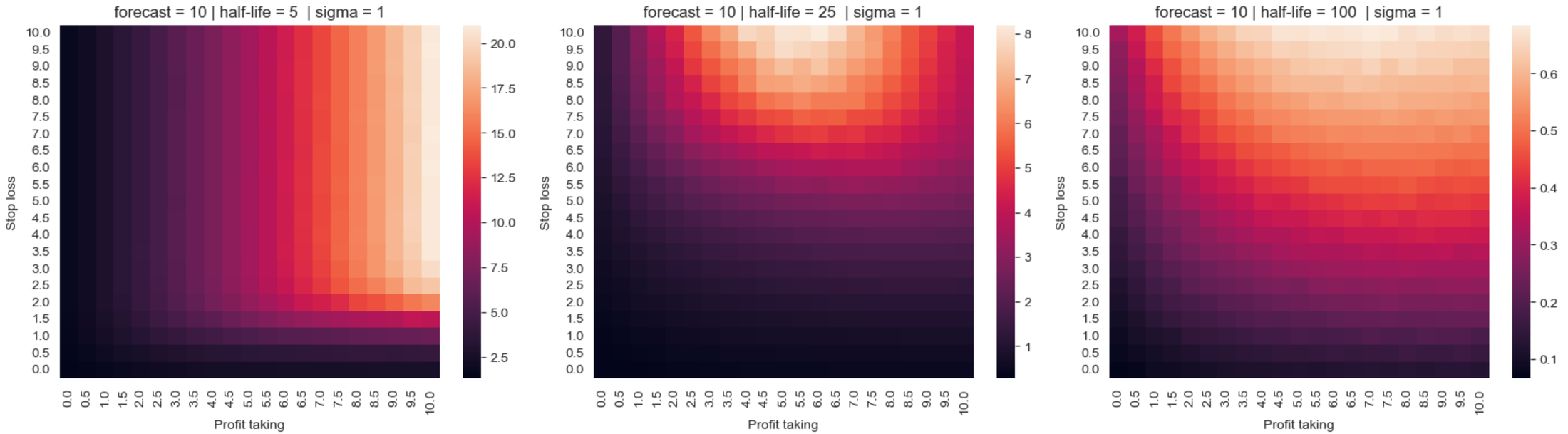
Monte Carlo simulation을 통한 Sharpe Ratio의 추정



- 장기 균형이 5 즉, 상승세를 보이는 시장의 경우, profit taking의 폭이 중간 정도이고, stopping loss의 폭이 클수록 sharpe ratio가 극대화 됨
- 상승장에서는 적당한 이익 실현과 손실시 보유하는 전략이 profit taking을 할 가능성이 높음

Backtest with Synthetic data

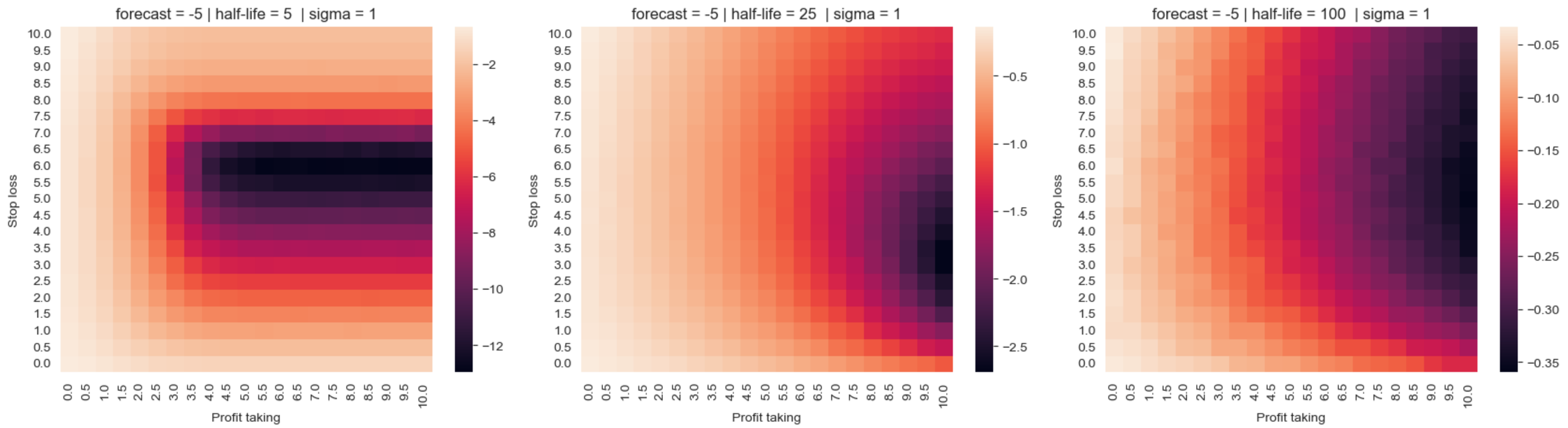
Monte Carlo simulation을 통한 Sharpe Ratio의 추정



- 장기 균형이 10 즉, 급격한 상승세를 보이는 시장의 경우, profit taking과 stopping loss의 폭이 모두 클수록 sharpe ratio가 극대화 됨
- 급격한 상승장에서는 길게 보유할수록 유리함

Backtest with Synthetic data

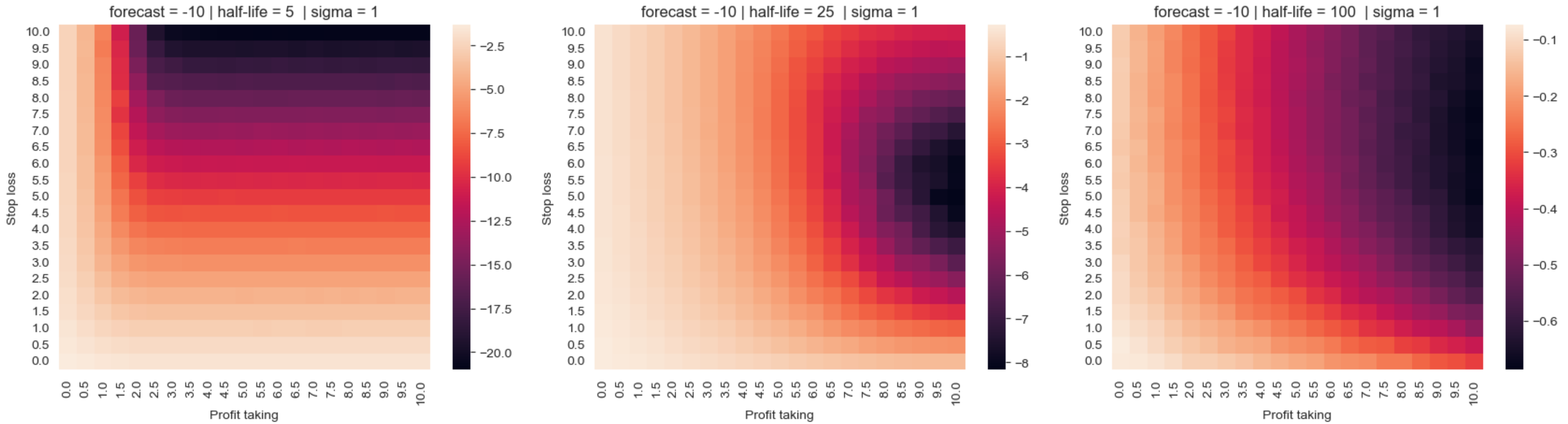
Monte Carlo simulation을 통한 Sharpe Ratio의 추정



- 장기 균형이 -5 즉, 하락세를 보이는 시장의 경우, profit taking과 stopping loss은 빠를수록 손실 폭을 최소화할 수 있음
- 하락장을 보이는 경우, 진입 자체가 문제가 되지만, 진입하게 되었을 경우 빠른 판단이 중요함

Backtest with Synthetic data

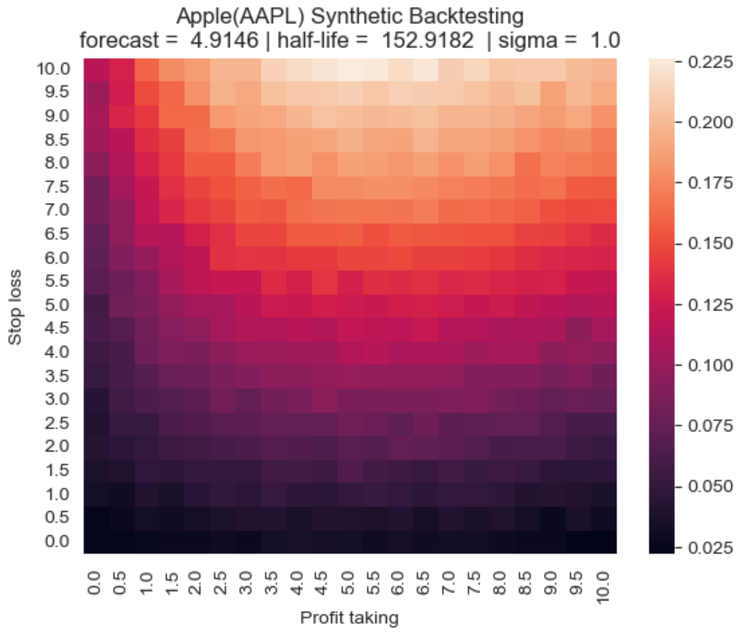
Monte Carlo simulation을 통한 Sharpe Ratio의 추정

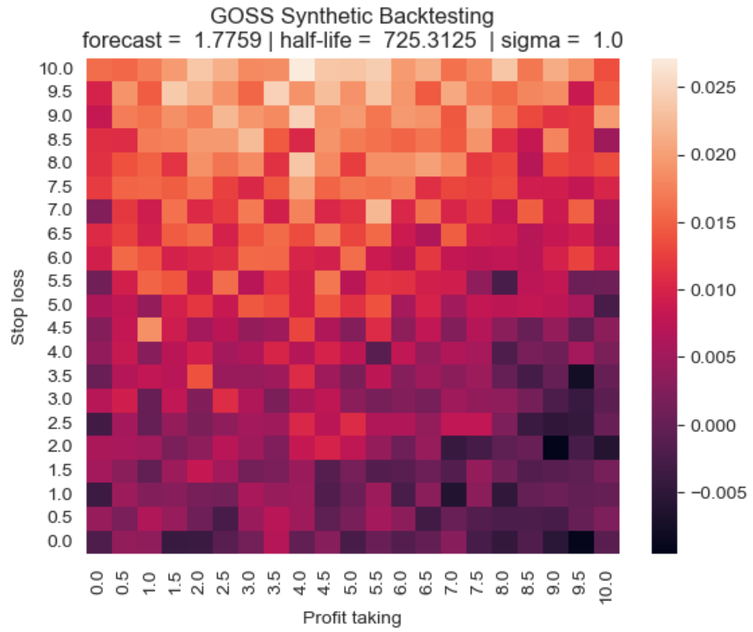


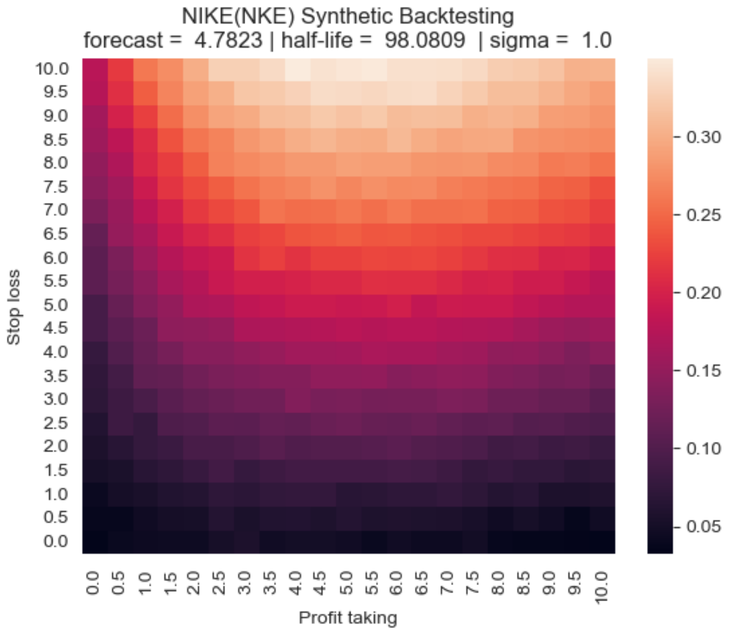
- 장기 균형이 -10 즉, 급격한 하락세를 보이는 시장의 경우, profit taking과 stopping loss 모두 빠를수록 손실 폭을 최소화할 수 있음
- 급격한 하락장을 보이는 경우, 이익실현과 손절이 모두 빨라야 피해를 최소화할 수 있음

Backtest with Synthetic data

Monte Carlo simulation을 통한 Sharpe Ratio의 추정







- Empirical data의 통계적 특성을 통한 Sharpe Ratio 관측행렬의 시각화
- 소형주인 GOSS의 경우, noise가 굉장히 많이 낀 것을 볼 수 있음. Half life 또한 매우 크게 나타난 것을 알 수 있음
- 대형주에 속하는 AAPL의 경우, 상승장인 특성을 알 수 있음. HI : 153

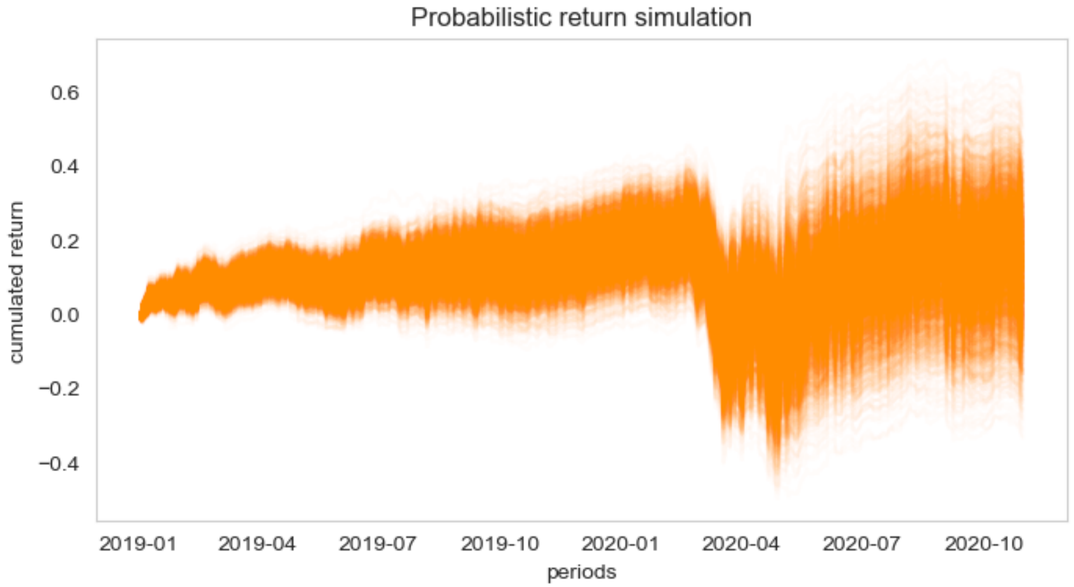
Deflated Sharpe Ratio

PSR로부터 DSR을 도출

- DSR은 SR에서 비정규 수익률이나 트랙 레코드 길이, 다중 테스트 혹은 선택에서 발생한 편향에 의한 부풀림 현상을 교정해 줌
- 표본의 임계감 0.05지점의 값을 사용

$$SR^* = \sqrt{V\left[\left\{\widehat{SR}_n\right\}\right]}\left((1-\gamma)Z^{-1}\left[1-\frac{1}{N}\right]+\gamma Z^{-1}\left[1-\frac{1}{N}e^{-1}\right]\right)$$

- 목표 SR을 도출



Conclusion

Monte Carlo Simulation의 중요성

- Monte Carlo simulation을 통해 다양한 확률 경로를 이용한 시뮬레이션이 가능
- 과거 확률 경로에 의존한 값이 아닌, 경험적 분포로부터 추정된 통계적 특성으로부터 neutral한 성과 추정 가능
- Ornstein Uhlenbeck process는 어디까지나 parametic method
- GAN의 Decoder를 이용해 Nonparametic stochastic process를 이용하면 좀 더 현실적인 성과 추정이 가능하지 않을까?
- 또는, Decoder를 통해 생성한 과정을 observation으로 두고, stochastic process를 통해 생성한 과정을 prior로 둔 뒤 posterior를 추정하여 생성하면, 좀 더 효율적인 Sharpe Ratio 추정이 가능하지 않을까?