Deep Learning Statistical Arbitrage

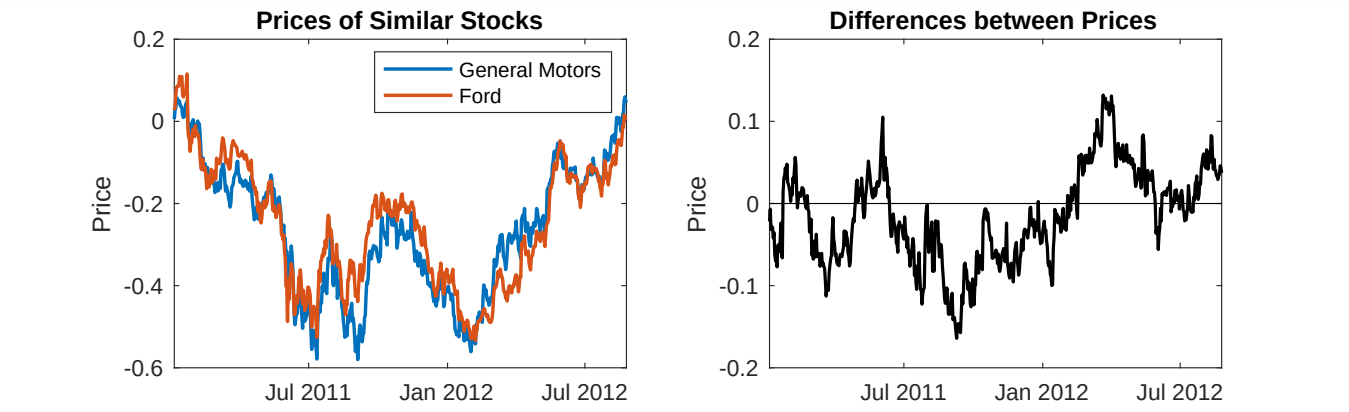
Jorge Guijarro-Ordonez, Markus Pelger, and Greg Zanotti

Stanford University

Motivation

Intuition: Pairs trading (simplest statistical arbitrage)

- Identify two "similar" stocks: e.g. GM and Ford
- Assumption: prices are on average similar
- Exploit temporal price differences between similar assets



Three components of statistical arbitrage:

1. Construct long-short portfolio identifying mispricing: $\epsilon_t = R_t^{\text{GM}} - R_t^{\text{Ford}}$
2. Extract trading signals by statistically modeling ϵ_t
3. Find optimal trading policy given signals: $\max \mathbb{E}\left[\text{payoff}_T\right]$

Fundamental Problem

Key elements of statistical arbitrage:

1. Arbitrage portfolios: How to generate long-short portfolios of similar assets?
2. Arbitrage signal: What are time-series patterns for temporary price deviations?
3. Arbitrage allocation: How to trade given the arbitrage signal?

Challenges:

1. Large number of assets with unknown similarities
2. Complex time-series patterns in price deviations
3. Optimal trading rules are complicated and depend on trading objective

Can machine learning help?

- Machine learning methods very flexible and deal with big data, but ...
- Important to set up the estimation problem correctly: Not a prediction problem!
- We use a trading objective function on residuals of asset pricing models

Key questions:

1. What is the "best solution" for the three key elements?
2. What matters for statistical arbitrage?
3. How much realistic arbitrage is in the market?

Contribution: Methodology

Our novel method: Deep learning statistical arbitrage

1. Statistical factor model including characteristics to get arbitrage portfolios
2. Convolutional neural network + Transformer to extract arbitrage signal:
   Flexible data driven time-series filter to learn complex time-series patterns
3. Neural network to map signals into allocations:
   Generalization of conventional "optimal stopping rules" for investment.

$\Rightarrow$ We integrate and optimize them for global economic objective:
Maximize risk-adjusted return under constraints.

$\Rightarrow$ Most advanced AI for NLP for time-series pattern detection

Novel conceptual framework:

- Provide unified framework to compare different statistical arbitrage methods:
  (1) portfolio generation, (2) signal extraction, (3) allocation decision
- Study each component and compare with conventional models
- Unifying time-series filter perspective for arbitrage signal

Contribution: Empirical

Comprehensive out-of-sample study on U.S. equities

- Daily returns for 19 years of 500 largest liquid stocks
- Consider most important risk factor models
- Comparisons include parametric and non-parametric mean-reversion models

Excellent out-of-sample performance:

- Empirically substantially outperforms all benchmark approaches out-of-sample
- Our arbitrage strategies achieve annual Sharpe ratios 4
- Annual returns of around 20% with less than 6% volatility
- Uncorrelated with conventional risk factors and market movements
- Survives realistic transaction and holding costs
- Stable over time and robust to tuning parameters

What matters for arbitrage trading?

- Robust to risk factors to identify similar assets
- Most important is time-series signal; flexible allocation model insufficient
- 4x better than parametric models, 2x better than non-parametric
- Global objective: extract time-series model for trading

Insight into the structure of arbitrage trading:

- "Smooth" trend and mean-reversion patterns
- Asymmetric policies: fast reaction on downtrends, cautious trading on uptrends

Literature (partial list)

Classical approaches to statistical arbitrage (parametric models)

- PCA + mean-reversion: Avellaneda and Lee (2010), Yeo and Papanicolaou (2017)
- Cointegration: Rad, Low and Faff (2016), Vidyamurthy (2004)
- Stochastic control: Cartea and Jaimungal (2016), Leung and Li (2015)
- Simple pairs trading: Gatev, Goetzmann and Rouwenhorst (2006)
- Intractable parametric models with ML: Mulvey, Sun, Wang, and Ye (2020)

Machine learning for asset pricing (explain risk premium not arbitrage)

- Pricing kernel: Chen, Pelger, Zhu (2019), Bryzgalova, Pelger, and Zhu (2019)
- Return prediction: Gu, Kelly and Xiu (2020),
- Factor models: Lettau and Pelger (2020), Kelly, Pruitt and Su (2019)

Machine learning for time-series (no trading objective)

- Time-series prediction: Lim and Zohren (2020), Krauss, Doa, and Huck (2017).

Model

Arbitrage portfolios

Excess returns of stocks follow a conditional factor model:

$$R_{n,t} = \beta_{n,t-1}^{\top} F_t + \epsilon_{n,t} \qquad t = 1, ..., T \text{ and } n = 1, ..., N_t$$

- K factors F_t capture systematic risk.
- Loadings $\beta_{t-1} \in \mathbb{R}^{N_t \times K}$ are general function of information at time $t-1$.

Factor models identify similar assets by similar exposures to risk factors

- Define *arbitrage portfolio* as residual portfolios:

$$\epsilon_{n,t} = R_{n,t} - \beta_{n,t-1}^{\top} F_t$$

- Arbitrage portfolios are only weakly cross-sectionally dependent.
- Arbitrage Pricing Theory implies $\mathbb{E}[\epsilon_{n,t}] = 0$.
- $\beta_{n,t-1}^{\top} F_t$ is "fair price" of $R_{n,t}$ and $\epsilon_{n,t}$ captures temporary mispricing

Arbitrage portfolios

Residuals with the empirically most important families of factor models:

1. Observed fundamental factors: Fama-French factors.
2. Statistical factors that explain correlations: PCA factors.
3. Conditional statistical factors where loadings are functions of firm characteristics: Instrumented PCA factors (Kelly, Pruitt and Su (2019)).

Factors are projections on returns without loss of generality:

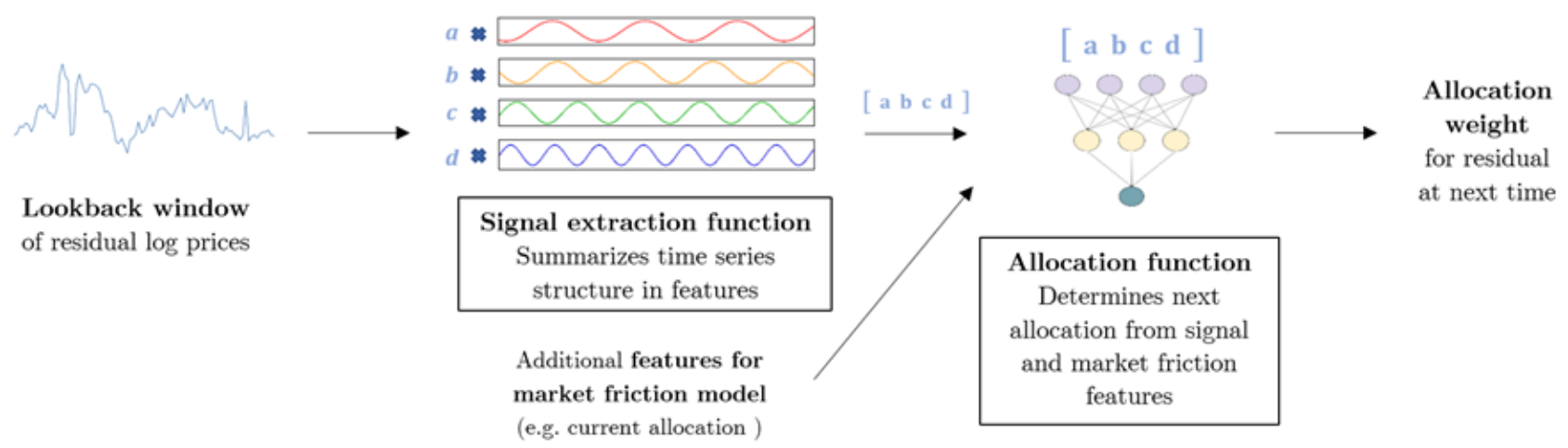
$$F_t = {w_{t-1}^F}^\top R_t.$$

Residuals are traded portfolios for factor implied matrix $\Phi_{t-1} \in \mathbb{R}^{N_t \times N_t}$:

$$\epsilon_t = R_t - \beta_{t-1}^T F_t = R_t - \beta_{t-1}^T w_{t-1}^F R_t = \underbrace{\left(I_{N_t} - \beta_{t-1}^T w_{t-1}^F\right)}_{\Phi_{t-1}} R_t.$$

$\Rightarrow$ Arbitrage portfolios are traded, factor-neutral, weakly correlated and mean-reverting portfolios of all stocks.

Arbitrage Signal and Allocation



Arbitrage trading has 2 steps given a cumulative residual

$$x := \epsilon_t^L := \left(\epsilon_{n,t-L} \quad \sum_{l=1}^{2} \epsilon_{n,t-L-1+l} \quad \cdots \quad \sum_{l=1}^{L} \epsilon_{n,t-L-1+l}\right)$$

1. The arbitrage signal function

$$\boldsymbol{\theta} \in \boldsymbol{\Theta} : \epsilon_{n,t-1}^L \mapsto \theta_{n,t-1}$$

   models the time series structure using last L cumulative residuals and estimates a sufficient statistic for trading.

2. The arbitrage allocation function

$$\boldsymbol{w}^{\epsilon} \in \boldsymbol{W} : \theta_{n,t-1} \mapsto w_{n,t-1}^{\epsilon}.$$

   assigns investment weights on residuals using only the estimated signal.

Estimation Problem

Estimation: For a given class of models maximize risk-adjusted return:

$$\max_{\boldsymbol{w}^{\epsilon} \in \boldsymbol{W}, \boldsymbol{\theta} \in \boldsymbol{\Theta}} \frac{\mathbb{E}\left[{w_{t-1}^{R}}^{\top} R_t \right]}{\sqrt{\mathrm{Var}({w_{t-1}^{R}}^{\top} R_t)}}$$

$$\text{s.t.} \qquad w_{t-1}^{R} = \frac{{w_{t-1}^{\epsilon}}^{\top} \Phi_{t-1}}{\| {w_{t-1}^{\epsilon}}^{\top} \Phi_{t-1} \|_1} \qquad \text{and} \qquad w_{t-1}^{\epsilon} = \boldsymbol{w}^{\epsilon}(\boldsymbol{\theta}(\epsilon_{t-1}^{L})).$$

- Main objective: Sharpe ratio, but we also consider mean-variance objective
- Extension includes trading costs
- Stock weights w_{t-1}^{R} add up to 1 $\Rightarrow$ implicit leverage constraint
- Many models have separate objective for signal estimation

We consider 3 key model classes for signal $\boldsymbol{\theta}$ and allocation $\boldsymbol{w}^{\epsilon}$:

1. Parametric model: mean-reversion model with thresholding rule
2. Pre-specified time-series filters and non-parametric allocation
3. Deep-learning arbitrage: data-driven time-series filter and allocation

$\Rightarrow$ We show what are the key elements for profitable arbitrage

First class: Parametric models

Classical mean reversion trading: (Avellaneda and Lee (2010))

- Each residual is modeled as an Ornstein-Uhlenbeck (OU) process

$$dX_t = \kappa(\mu - X_t)dt + \sigma dB_t$$

- The allocation is a threshold rule on the ratio $\frac{X_t - \mu}{\sigma/\sqrt{2\kappa}}$.
- In our framework, this corresponds to

$$\theta^{\text{OU}}(x) = (\hat{\kappa}, \hat{\mu}, \hat{\sigma}, x_L), \quad \boldsymbol{w}^X\left(\theta^{\text{OU}}\right) = \begin{cases} -1, & \text{if } \frac{x_L - \hat{\mu}}{\hat{\sigma}/\sqrt{2\hat{\kappa}}} > c_{\text{thres}} \\ 1 & \text{if } \frac{x_L - \hat{\mu}}{\hat{\sigma}/\sqrt{2\hat{\kappa}}} < -c_{\text{thres}} \\ 0 & \text{otherwise} \end{cases}$$

  where c_{thres} is chosen optimally.

Limitations: Parametric model might be misspecified (eg. trends, multiple mean reversion frequencies, etc.), restrictive allocation function.

Second class: Pre-specified filter with neural network

Signal θ: General time-series model

- Pre-specified linear filter $\theta_l = \sum_{j=1}^{L} W_j^{\text{filter}} x_j$ (given matrix $W^{\text{filter}} \in \mathbb{R}^{L \times L}$)
- Includes ARMA models, discretized OU, etc.
- Frequency filters are the most relevant filters for mean reversion patterns:
- We use Fast Fourier Transform (FFT) for a frequency decomposition:

$$x_l = a_0 + \sum_{j=1}^{L/2-1} \left(a_j \cdot \cos\left(\frac{2\pi j}{L} l\right) + b_j \cdot \sin\left(\frac{2\pi j}{L} l\right) \right) + a_{L/2} \cos(\pi l).$$

- Signal are the "loadings" on long and short-term reversal patterns:

$$\boldsymbol{\theta}^{\text{FFT}}(x) = (a_0, \ldots, a_{L/2}, b_1, \ldots b_{L/2-1})$$

Allocation w^{ϵ}: Flexible non-parameteric function with regularization

- $\boldsymbol{g}^{\text{FFN}}$ is estimated with feedforward neural network (FFN)

$$\boldsymbol{w}^{\epsilon|\text{FFT}}\left(\theta^{\text{FFT}}\right) = \boldsymbol{g}^{\text{FFN}}\left(\theta^{\text{FFT}}\right).$$

Limitation: Choice of pre-specified filter limits the time-series patterns.

Third class: Convolutional Network with Transformer

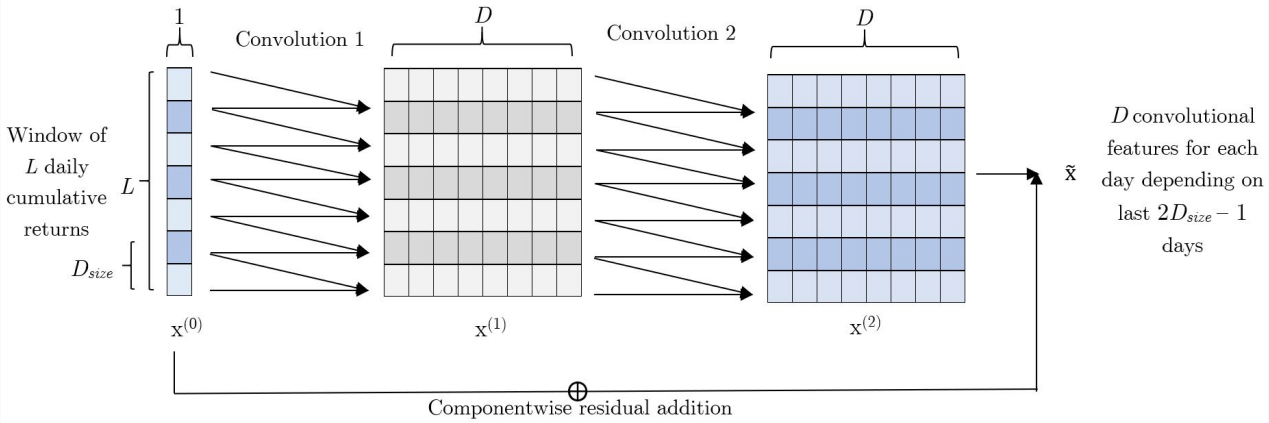
Our novel model: Data driven time-series filter based on most advanced deep learning tools for pattern detection

- Convolutional neural networks (CNN) are data-driven non-linear local filters
- Transformers learn global dependency patterns between local filters
- CNN+Transformer is a flexible non-linear filter that can learn any time-series pattern
- Examples of global "pattern factors"
  - Mean-reversion: cyclical combination of local curvature patterns
  - Trend: Monotonic combination of local drifts
- Signal $\boldsymbol{\theta}^{\text{CNN+Trans}}(x)$ is the "exposure" to pattern factors
- Allocation function $\boldsymbol{w}^{\epsilon}$ is a flexible FFN:

$$\boldsymbol{w}^{\epsilon|\text{CNN+Trans}}\left(\theta^{\text{CNN+Trans}}\right) = \boldsymbol{g}^{\text{FFN}}\left(\theta^{\text{CNN+Trans}}\right).$$

- Joint estimation of signal and allocation function with trading objective

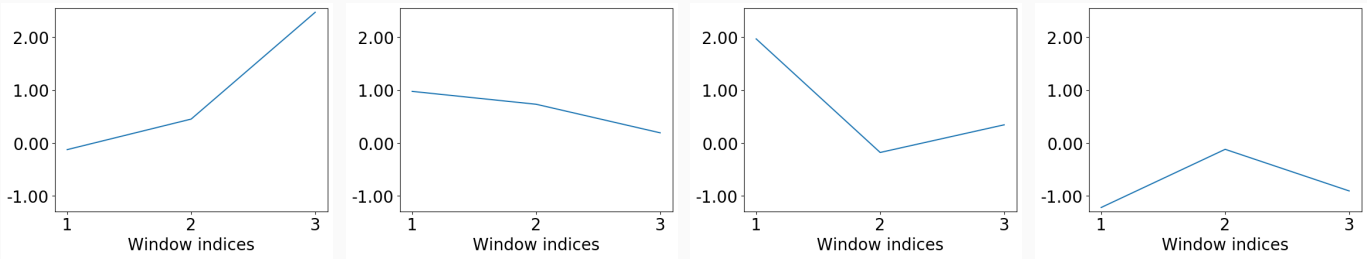
Convolutional Network Intuition



- The network applies to residual time series ϵ_t^L a combination of local estimated linear filters W^{local} followed by non-linear transformations:

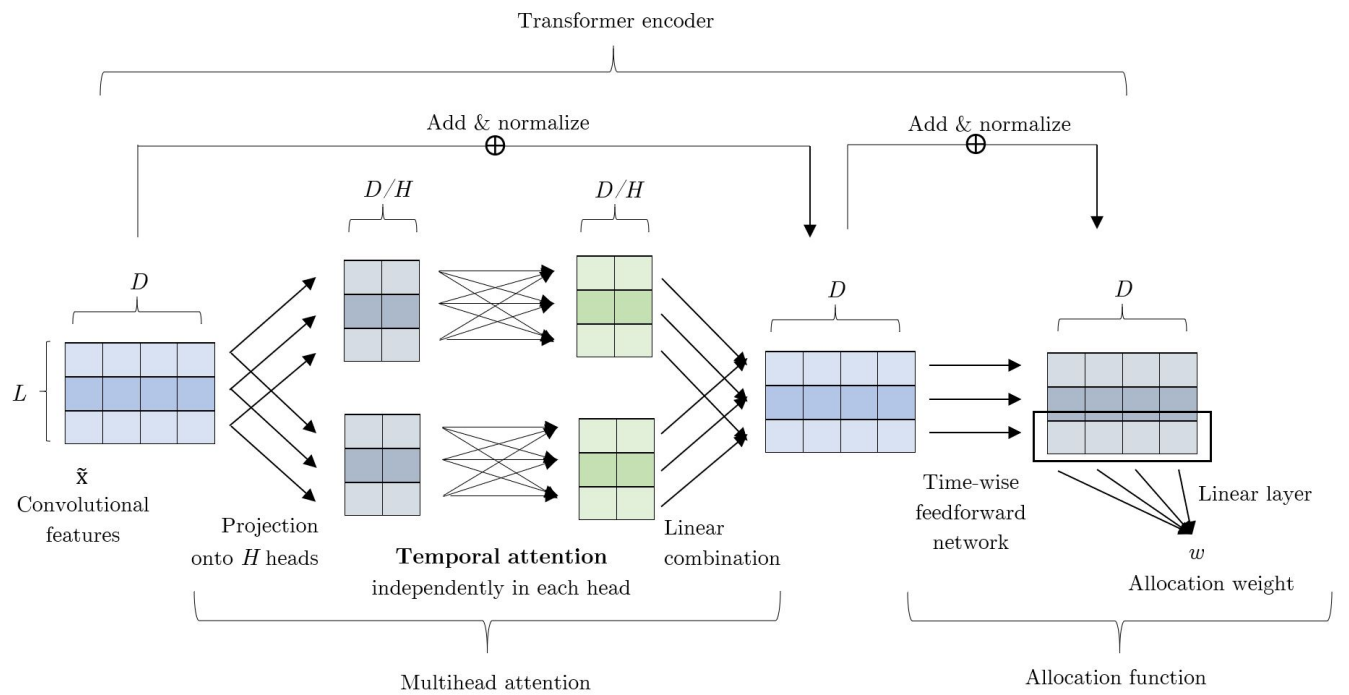
$$y_l^{(0)} = \sum_{m=1}^{D_{\text{size}}} W_m^{\text{local}} x$$

$\Rightarrow$ Represent time-series $x \in \mathbb{R}^L$ in terms of D local patterns $\tilde{x} \in \mathbb{R}^{L \times D}$



(a) Upward trend (b) Downward trend (c) Up reversal (d) Down reversal

Transformer Network Intuition



Transformer captures temporal dependencies between local patterns

$$h_i = \sum_{l=1}^{L} \alpha_{i,l} \tilde{x}_l \qquad \text{with } \alpha_{i,l} = \boldsymbol{\alpha}_i (\tilde{x}_L, \tilde{x}_l) \text{ for } l = 1, ..., L \text{ and } i = 1, .., H$$

- H global patterns specified by "attention weights" $\alpha_i \in \mathbb{R}^L$.
- Attention heads h_i are "loadings" for a specific "pattern factor" α_i
- Transformer estimate flexible attention weight functions

Empirical Analysis

Data

Out-of-sample analysis on U.S. equity data:

- 19 years of large cap U.S. daily stock returns from Jan 1998 to Dec 2016
- Only stocks with prior month market cap $> 0.01\%$ of total market cap
  $\Rightarrow \sim 550$ large cap stocks/month $\approx$ S&P 500
  Most liquid stocks to avoid trading frictions
- For IPCA, supplement with 46 monthly firm characteristics for each stock and month (starting in 1978).

Implementation:

- All results are out-of-sample
- We use $L = 30$ days lookback windows of returns as input for signal.
- We retrain functions every half year using rolling windows of 4 years.
- Factors models are estimated OOS daily on rolling window of 60 days
- Main analysis with Sharpe ratio objective

Arbitrage Portfolios

Residuals with the empirically most important families of factor models:

1. Fama-French factors for 1, 3, 5, 8 factors.
   market, size, value, investment, profitability, momentum, short-term and long-term reversal
2. PCA factors for 1, 3, 5, 8, 10, 15 factors.
3. IPCA model of Kelly, Pruitt, and Su (2019), for 1, 3, 5, 8, 10, 15 factors.
4. "0-factor model": original stocks instead of residuals.

Given the residuals, we estimate arbitrage signals and allocations for

1. Ornstein-Uhlenbeck estimation with threshold rule (OU+Thres).
2. Fast Fourier Transform with feedforward network (Fourier+FFN).
3. Convolutional network with transformer (CNN+Trans).

and, for completeness,

4. OU estimation with feedforward network (OU+FFN).
5. Just a feedforward network without time-series filter (FFN)

OOS Annualized Performance

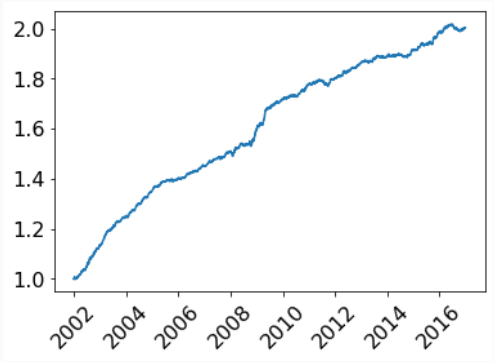
	Factors	Fama-French			PCA			IPCA		
Model	K	SR	μ	σ	SR	μ	σ	SR	μ	σ
CNN+	0	1.64	13.7%	8.4%	1.64	13.7%	8.4%	1.64	13.7%	8.4%
Trans	5	3.21	4.6%	1.4%	3.36	14.3%	4.2%	4.16	8.7%	2.1%
FFT+	0	0.36	4.9%	13.6%	0.36	4.9%	13.6%	0.36	4.9%	13.6%
FFN	5	1.66	3.1%	1.8%	1.98	12.4%	6.3%	1.90	7.7%	4.1%
OU+	0	-0.18	-2.4%	13.3%	-0.18	-2.4%	13.3%	-0.18	-2.4%	13.3%
Thres	5	0.38	0.9%	2.3%	0.73	4.4%	6.1%	0.97	3.8%	4.0%

- Arbitrage trading has to be applied to residuals and not returns
- Results do not substantially improve after regressing out 5 factors
- CNN+Transformer strongly dominates all models
- Average return μ is high in spite of leverage constraint
- Arbitrage trading qualitatively robust to choice of factor model
- Fourier+FFN only 50% of CNN+Trans $\Rightarrow$ flexible time-series filter crucial!
- Conventional OU+Thres only 25% of CNN+Trans
  $\Rightarrow$ Too restrictive model!

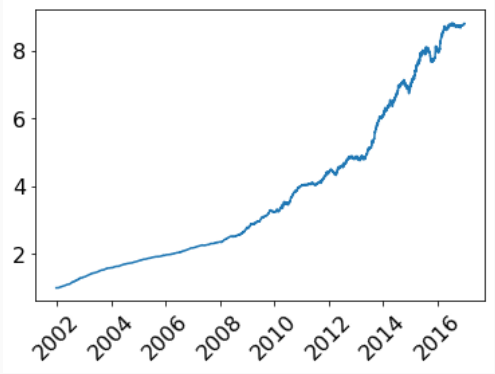
OOS Annualized Performance

	Factors	Fama-French			PCA			IPCA		
Model	K	SR	μ	σ	SR	μ	σ	SR	μ	σ
CNN + Trans	0	1.64	13.7%	8.4%	1.64	13.7%	8.4%	1.64	13.7%	8.4%
	1	3.68	7.2%	2.0%	2.74	15.2%	5.5%	3.22	8.7%	2.7%
	3	3.13	5.5%	1.8%	3.56	16.0%	4.5%	3.93	8.6%	2.2%
	5	3.21	4.6%	1.4%	3.36	14.3%	4.2%	4.16	8.7%	2.1%
	8	2.49	3.4%	1.4%	3.02	12.2%	4.0%	3.95	8.2%	2.1%
	10	-	-	-	2.81	10.7%	3.8%	3.97	8.0%	2.0%
	15	-	-	-	2.30	7.6%	3.3%	4.17	8.4%	2.0%
Fourier + FFN	0	0.36	4.9%	13.6%	0.36	4.9%	13.6%	0.36	4.9%	13.6%
	1	0.89	3.2%	3.5%	0.80	8.4%	10.6%	1.24	6.3%	5.0%
	3	1.32	3.5%	2.7%	1.66	11.2%	6.7%	1.77	7.8%	4.4%
	5	1.66	3.1%	1.8%	1.98	12.4%	6.3%	1.90	7.7%	4.1%
	8	1.90	3.1%	1.6%	1.95	10.1%	5.2%	1.94	7.8%	4.0%
	10	-	-	-	1.71	8.2%	4.8%	1.93	7.6%	3.9%
	15	-	-	-	1.14	4.8%	4.2%	2.06	7.9%	3.8%
OU + Thresh	0	-0.18	-2.4%	13.3%	-0.18	-2.4%	13.3%	-0.18	-2.4%	13.3%
	1	0.16	0.6%	3.8%	0.21	2.1%	10.4%	0.60	3.0%	5.1%
	3	0.54	1.6%	3.0%	0.77	5.2%	6.8%	0.88	3.8%	4.3%
	5	0.38	0.9%	2.3%	0.73	4.4%	6.1%	0.97	3.8%	4.0%
	8	1.16	2.8%	2.4%	0.87	4.4%	5.1%	0.91	3.5%	3.8%
	10	-	-	-	0.63	2.9%	4.6%	0.86	3.1%	3.6%
	15	-	-	-	0.62	2.4%	3.8%	0.93	3.2%	3.5%

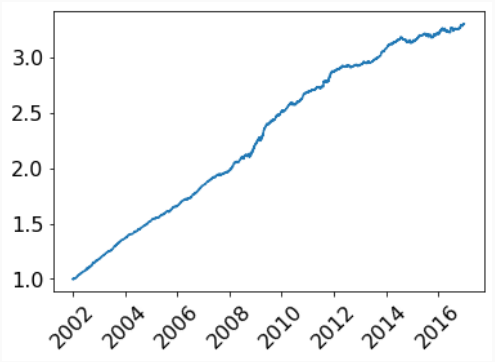
Cumulative OOS Returns of Different Arbitrage Strategies



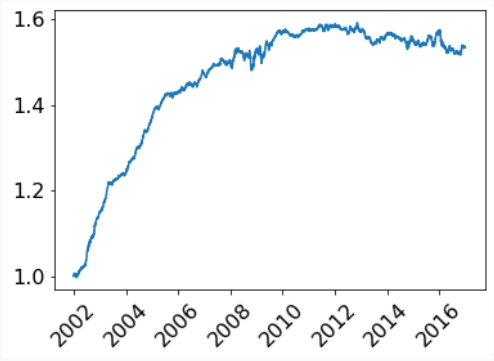
(a) CNN+Trans, Fama-French 5



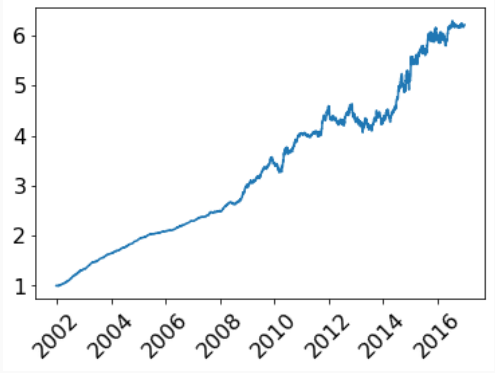
(b) CNN+Trans, PCA 5



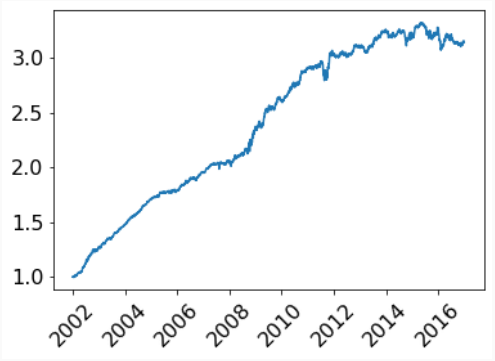
(c) CNN+Trans, IPCA 5



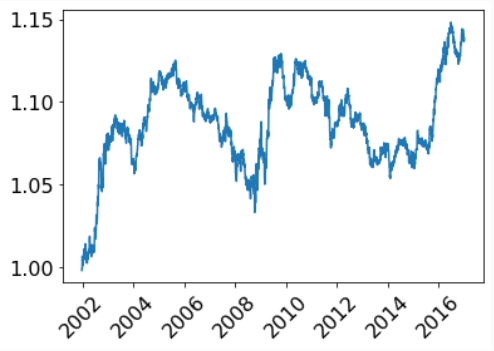
(d) FFT+FFN, Fama-French 5



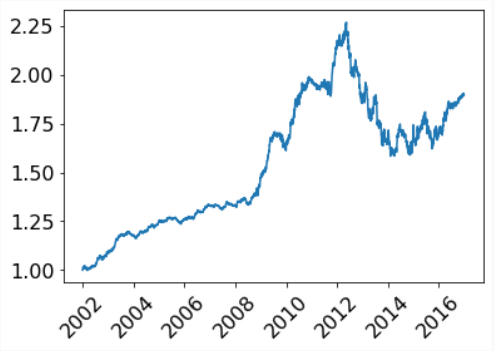
(e) FFT+FFN, PCA 5



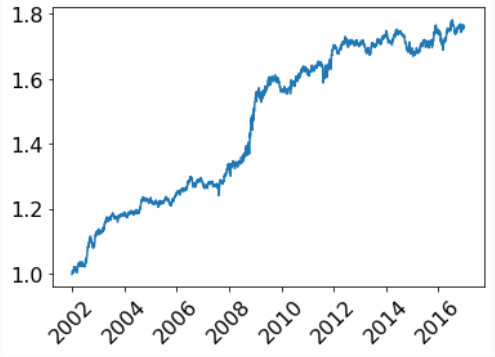
(f) FFT+FFN, IPCA 5



(g) OU+Thresh Fama-French 5



(h) OU+Thresh PCA 5



(i) OU+Thresh IPCA 5

Significance of Arbitrage Alphas

CNN+Trans model						
	Fama-French		PCA		IPCA	
K	0	5	0	5	0	5
α	11.6%	4.5%	11.6%	14.1%	11.6%	8.3%
μ	13.7%	4.6%	13.7%	14.3%	13.7%	8.7%
t_α	6.4^{***}	12^{***}	6.4^{***}	13^{***}	6.4^{***}	16^{***}
t_μ	6.3^{***}	12^{***}	6.3^{***}	13^{***}	6.3^{***}	16^{***}
R^2	30.3%	2.3%	30.3%	1.3%	30.3%	3.9%

- Time-series regression of 8 asset pricing factors:
  Fama-French 5 + momentum, short-term and long-term reversal factors
  Pricing errors α, t-statistics t_α and R^2 of regression
- Mean μ and corresponding t-statistics t_μ of arbitrage stratgies
- CNN+Transformer arbitrage is statistically significant and not subsumed by conventional risk factors
- Arbitrage strategies orthogonal to conventional risk factors
- Conventional mean-reversion trading explained by conventional risk factors.

Significance of Arbitrage Alphas

CNN+Trans model																
	Fama-French					PCA					IPCA				
K	α	t_α	R^2	μ	t_μ	α	t_α	R^2	μ	t_μ	α	t_α	R^2	μ	t_μ
0	11.6%	6.4^{***}	30.3%	13.7%	6.3^{***}	11.6%	6.4^{***}	30.3%	13.7%	6.3^{***}	11.6%	6.4^{***}	30.3%	13.7%	6.3^{***}
1	7.0%	14^{***}	2.4%	7.2%	14^{***}	14.9%	10^{***}	0.6%	15.2%	11^{***}	8.1%	12^{***}	9.5%	8.7%	12^{***}
3	5.5%	12^{***}	1.2%	5.5%	12^{***}	15.8%	14^{***}	1.7%	16.0%	14^{***}	8.2%	15^{***}	6.0%	8.6%	15^{***}
5	4.5%	12^{***}	2.3%	4.6%	12^{***}	14.1%	13^{***}	1.3%	14.3%	13^{***}	8.3%	16^{***}	3.9%	8.7%	16^{***}
8	3.3%	9.4^{***}	2.1%	3.4%	9.6^{***}	12.0%	12^{***}	0.9%	12.2%	12^{***}	7.8%	15^{***}	5.0%	8.2%	15^{***}
10	-	-	-	-	-	10.5%	11^{***}	0.7%	10.7%	11^{***}	7.7%	15^{***}	4.0%	8.0%	15^{***}
15	-	-	-	-	-	7.5%	8.8^{***}	0.5%	7.6%	8.9^{***}	8.1%	16^{***}	4.2%	8.4%	16^{***}

- Time-series regression of 8 asset pricing factors:
  Fama-French 5 + momentum, short-term and long-term reversal factors
  Pricing errors α, t-statistics t_α and R^2 of regression
- Mean μ and corresponding t-statistics t_μ of arbitrage stratgies
- CNN+Transformer arbitrage is statistically significant and not subsumed by conventional risk factors
- Arbitrage strategies orthogonal to conventional risk factors
- Conventional mean-reversion trading explained by conventional risk factors.

Mean-Variance Objective

CNN+Trans model, mean-variance objective function

	Fama-French			PCA			IPCA		
K	SR	μ	σ	SR	μ	σ	SR	μ	σ
0	0.83	9.5%	11.4%	0.83	9.5%	11.4%	0.83	9.5%	11.4%
1	3.15	10.5%	3.3%	2.21	27.3%	12.3%	2.83	15.9%	5.6%
3	2.95	7.8%	2.6%	2.38	22.6%	9.5%	3.13	17.9%	5.7%
5	3.03	5.9%	2.0%	2.75	19.6%	7.1%	3.21	18.2%	5.7%
8	2.96	4.2%	1.4%	2.68	16.6%	6.2%	3.18	17.0%	5.4%
10	-	-	-	2.67	15.3%	5.7%	3.21	16.6%	5.2%
15	-	-	-	2.20	8.7%	4.0%	3.34	16.3%	4.9%

Alternative mean-variance objective function:

$$\max_{\boldsymbol{w}^{\epsilon} \in \boldsymbol{W}, \boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathbb{E}[{w_{t-1}^{R}}^{\top} R_t] - \gamma \mathrm{Var}({w_{t-1}^{R}}^{\top} R_t)$$

$$\text{s.t.} \qquad w_{t-1}^{R} = \frac{{w_{t-1}^{\epsilon}}^{\top} \Phi_{t-1}}{\| {w_{t-1}^{\epsilon}}^{\top} \Phi_{t-1} \|_1} \qquad \text{and} \qquad w_{t-1}^{\epsilon} = \boldsymbol{w}^{\epsilon}(\boldsymbol{\theta}(\epsilon_{t-1}^{L})).$$

- Increase mean return while maintaining leverage constraint of $\|w_{t-1}^{R}\| = 1$
- Here we set risk aversion to $\gamma = 1$
- Annual returns up to 20% while volatility is only half of market.
- Slightly lower Sharpe ratios

Importance of Time-Series Signal

	Factors	Fama-French			PCA			IPCA		
Model	K	SR	μ	σ	SR	μ	σ	SR	μ	σ
FFN	0	0.57	8.8%	15.3%	0.57	8.8%	15.3%	0.57	8.8%	15.3%
	1	0.60	2.0%	3.3%	0.53	6.2%	11.7%	1.07	6.5%	6.1%
	3	1.02	2.6%	2.6%	1.15	8.2%	7.2%	1.50	7.6%	5.0%
	5	1.32	2.3%	1.7%	1.42	9.8%	6.9%	1.55	7.3%	4.7%
	8	1.31	2.1%	1.6%	0.84	5.1%	6.1%	1.56	7.2%	4.6%
	10	-	-	-	0.70	3.5%	5.0%	1.48	7.0%	4.7%
	15	-	-	-	0.51	2.4%	4.8%	1.68	7.5%	4.5%

Is the a time-series signal function actually needed?

- Apply flexible FFN directly to residuals without time-series model
- Results are substantially worse than Fourier+FFN
- FFN is not efficient enough to learn complex dependencies with limited data

Additional Results

Stability over time:

- Results are robust to length of local rolling window
  Essentially identical results for $L = 60$ rolling lockback window
- Constant signal and allocation function capture most arbitrage information
  30% decrease of performance for constant model ($T_{train} = 4$ or 8 years)
  Constant CNN+Transformer still substantially outperforms re-estimated benchmark models

Robustness to tuning parameters:

- Results very robust to all tuning parameters
- General structure of the problem important, but not number of layers, etc.

Dependency between strategies:

- Between different factor models only weakly correlated (0.2 to 0.45)
- Within factor family model high correlation (0.4 to 0.85)

Unconditional means without allocation:

- Equally weighted residuals have mean returns $< 1\%$
- Need to apply signal and trading policy to residuals for profitable trading

Trading Frictions and Transaction Costs

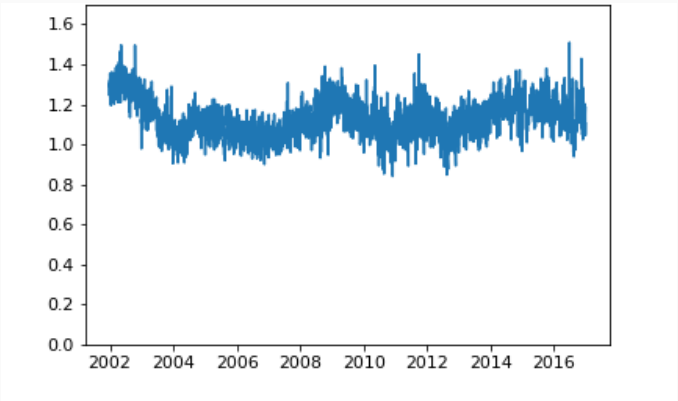
IPCA factor model						
	Sharpe ratio			Mean-variance		
K	SR	μ	σ	SR	μ	σ
0	0.52	8.5%	16.3%	0.22	2.6%	11.9%
1	0.85	5.9%	6.9%	0.86	5.5%	6.4%
3	1.24	6.6%	5.4%	1.16	6.9%	5.9%
5	1.11	5.5%	5.0%	1.02	5.3%	5.3%
10	0.98	5.1%	5.2%	1.04	5.4%	5.2%
15	0.94	4.8%	5.1%	1.02	5.1%	5.0%

- Include trading costs for high turnover and large short-selling positions:
  $\text{cost}(w^R_{t-1}, w^R_{t-2}) = 0.0005\|w^R_{t-1} - w^R_{t-2}\|_{L^1} + 0.0001\|\min(w^R_{t-1}, 0)\|_{L^1}$
  5 basis points per transaction and 1 basis point per short position
- No market impact as we only trade in the largest most liquid stocks
- Lower bound on profitability: less turnover with sparse factors, etc.

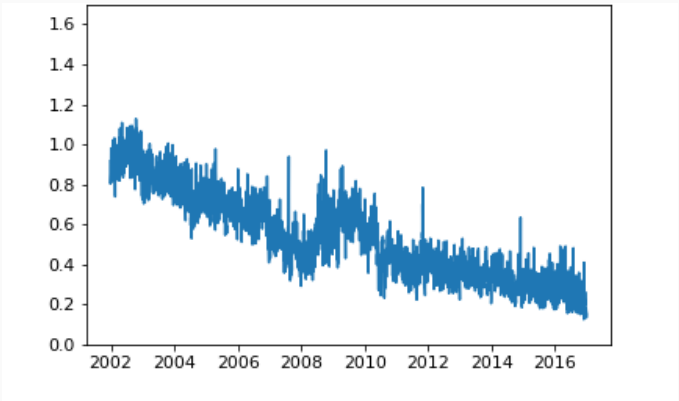
$\Rightarrow$ Arbitrage trading retains economic significance in presence of trading costs

Turnover and Short Selling

Turnover with and without trading friction objective:

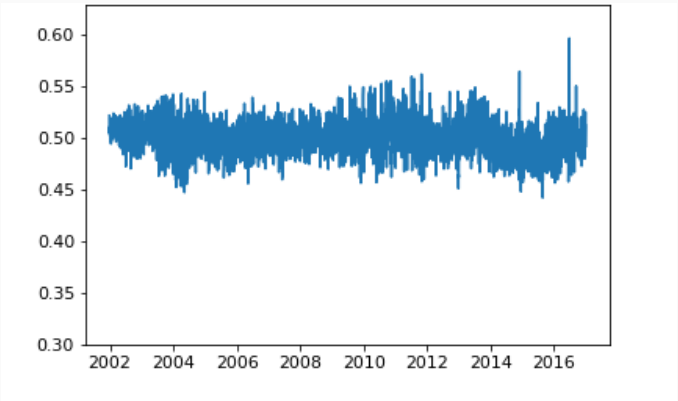


(a) No Trading Friction Objective

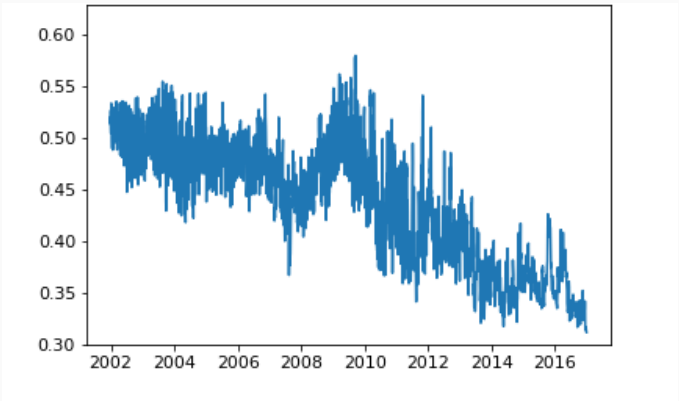


(b) With Trading Friction Objective

Proportion of short allocation weights:



(a) No Trading Friction Objective

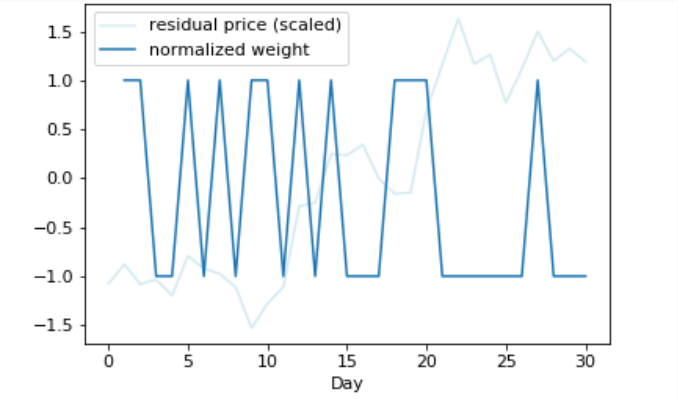


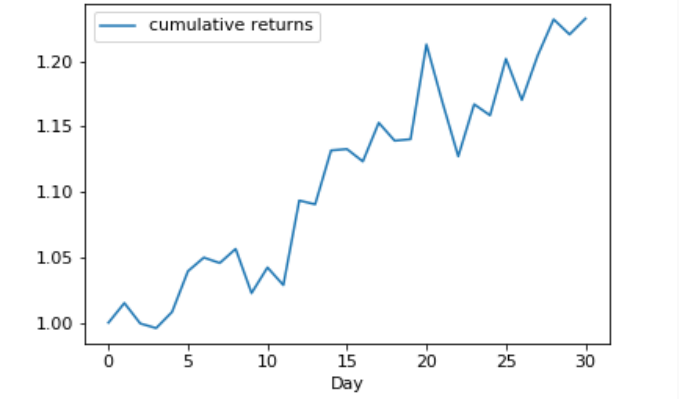
(b) With Trading Friction Objective

$\Rightarrow$ The effect of trading frictions is time-varying and our model can exploit particularly profitable arbitrage time periods by increasing trading and short positions.

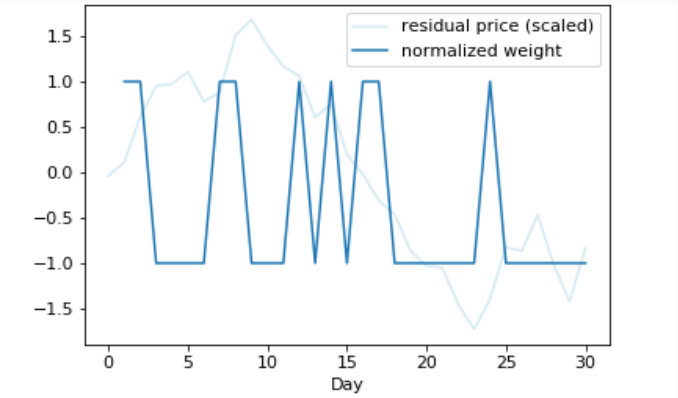
Estimated Structure: Dissecting the CNN+Transformer Model with IPCA-5

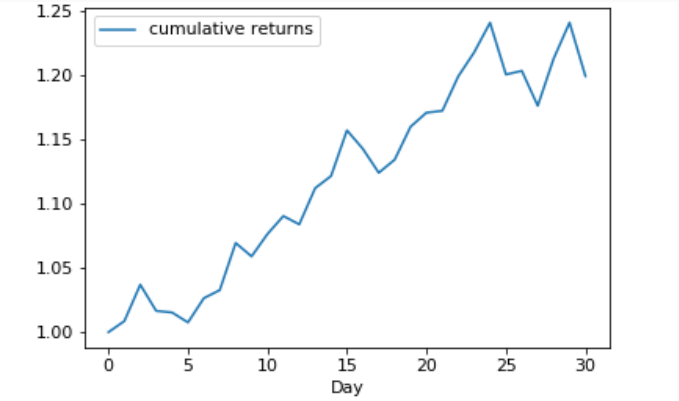
Examples of Allocation and Returns of CNN+Transformer Strategy





(a) Example 1: Mean-reversion



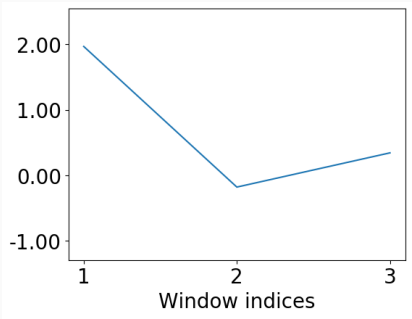


(b) Example 2: Trend

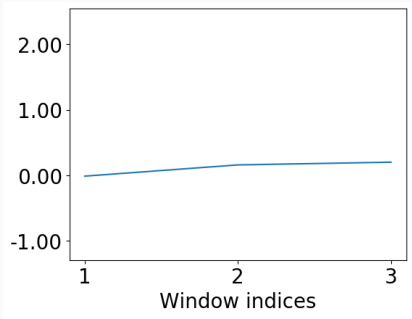
Sample of representative residuals with out-of-sample arbitrage trading

- Positive allocations for positive changes and vice-a-versa
- CNN detects global and local trend and reversion patterns

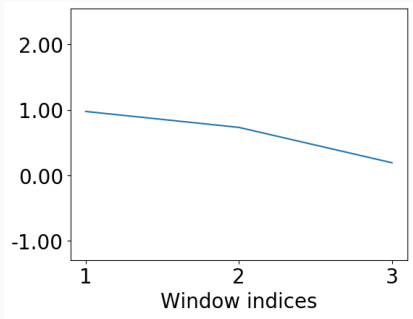
CNN: Local Basic Patterns



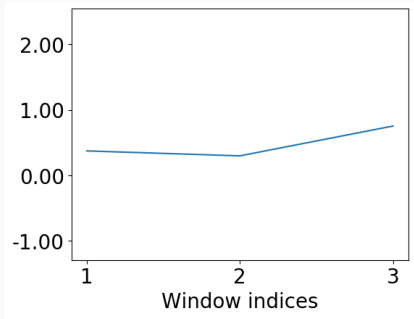
(a) Basic pattern 1



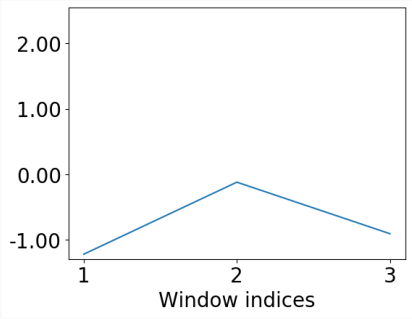
(b) Basic pattern 2



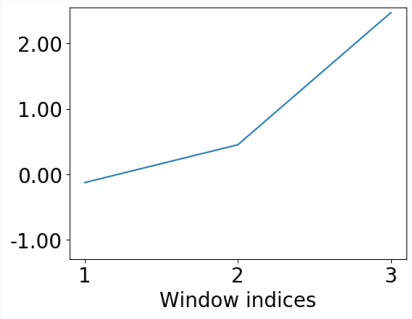
(c) Basic pattern 3



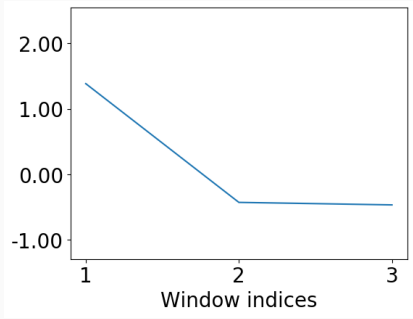
(d) Basic pattern 4



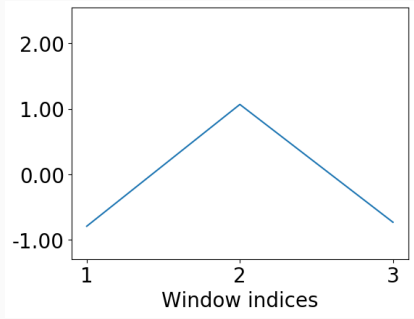
(e) Basic pattern 5



(f) Basic pattern 6



(g) Basic pattern 7

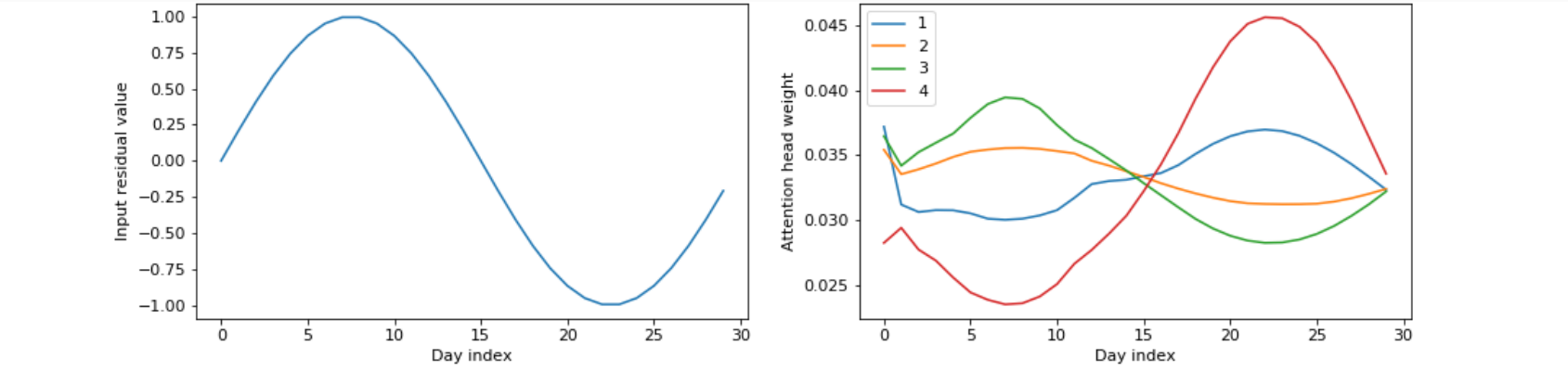


(h) Basic pattern 8

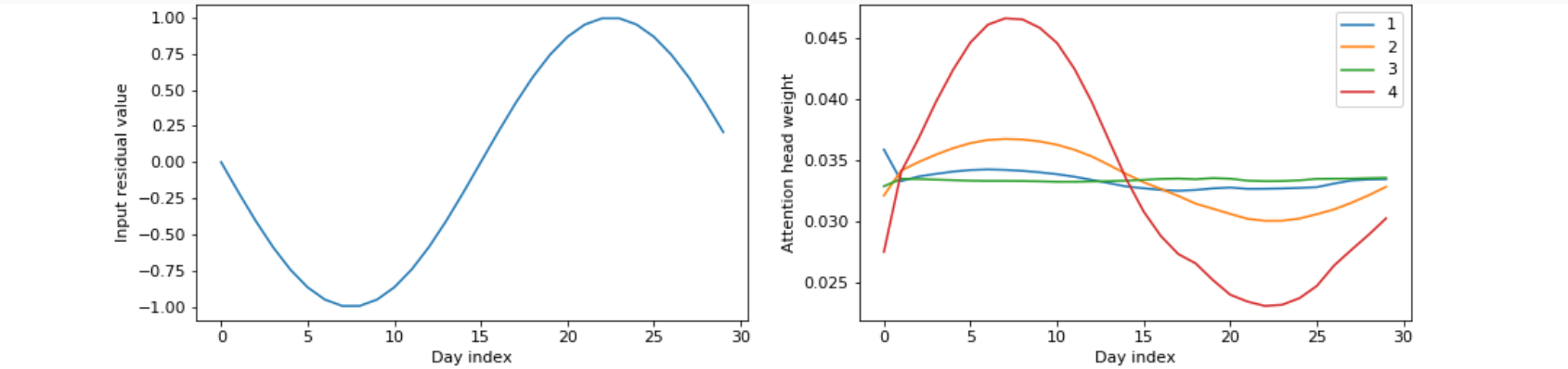
Local filters estimated by CNN to capture relative local patterns

- Basic patterns are "building blocks" for the global patterns
- Visualizations of non-linear 3-dimensional local filters into orthogonal two-dimensional local linear filters
- Sufficient to construct any smooth trend and mean-reversion patterns.

Example Attention Weights for Sinusoidal Residual Inputs



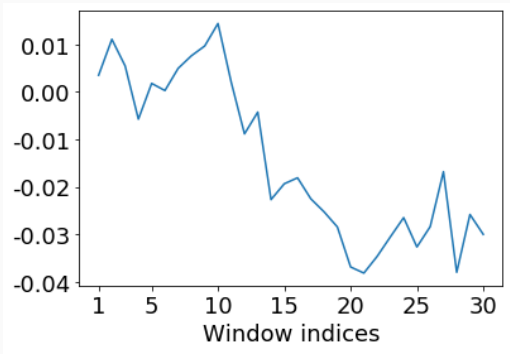
(a) Input residual and attention head weights for $x_l = \sin\left(2\pi \frac{l}{30}\right)$



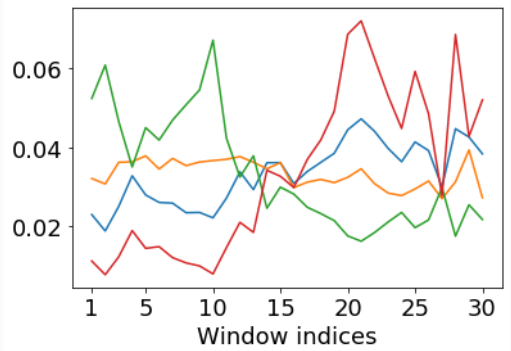
(b) Input residual and attention head weights for for $x_l = \sin\left(2\pi \frac{l+15}{30}\right)$

- Attention head weight 4: negative reversal factor
- Attention head weight 3: early reversal factor
- Attention head weight 1: low-frequency downturn factor

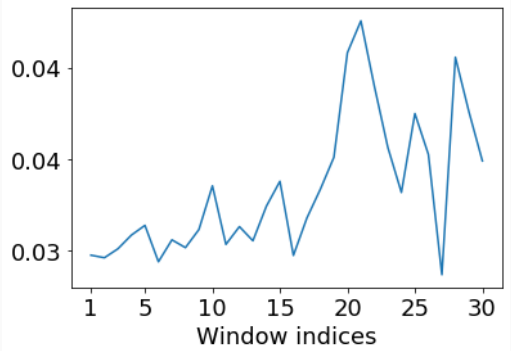
CNN+Transformer Model Structure for Representative Residual



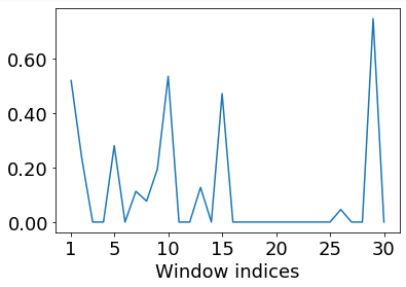
(a) Cumulative residual



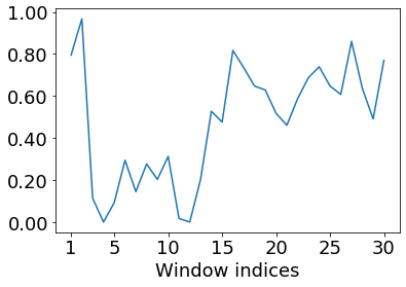
(b) Attention weights per head



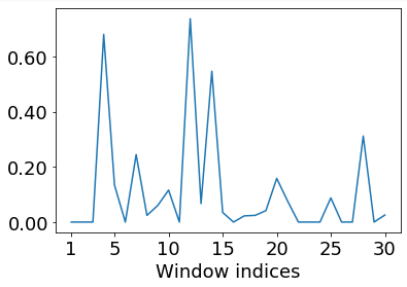
(c) Average attention weights



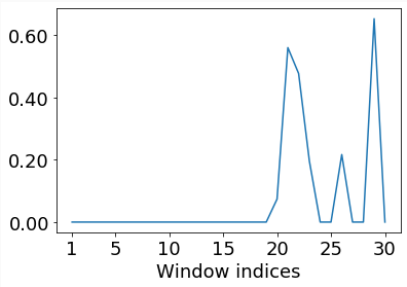
(d) 1st CNN activation



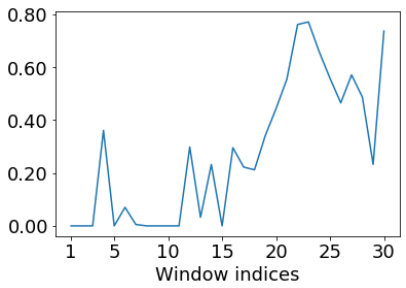
(e) 2nd CNN activation



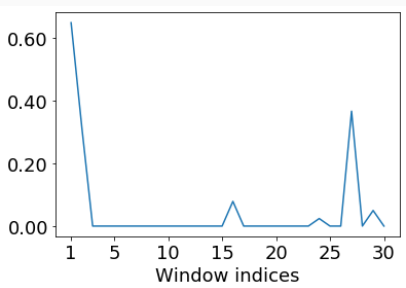
(f) 3rd CNN activation



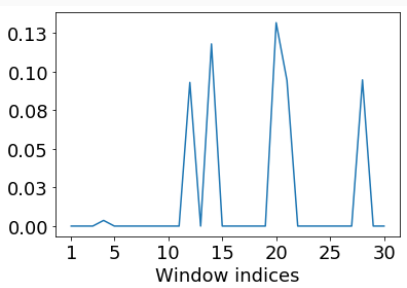
(g) 4th CNN activation



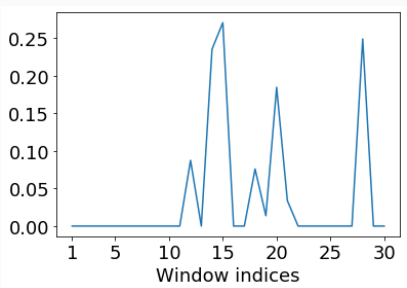
(h) 5th CNN activation



(i) 6th CNN activation

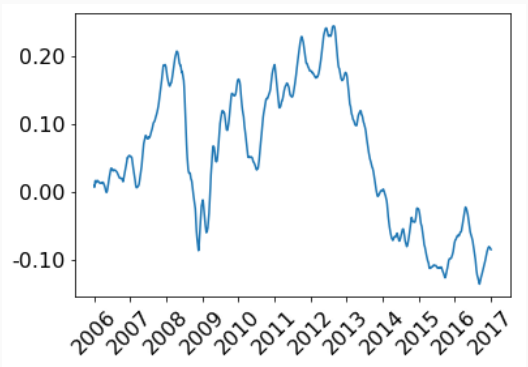


(j) 7th CNN activation

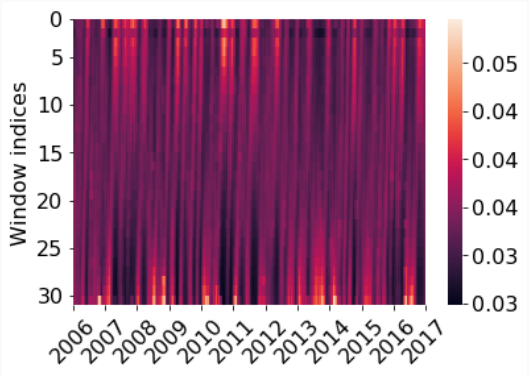


(k) 8th CNN activation

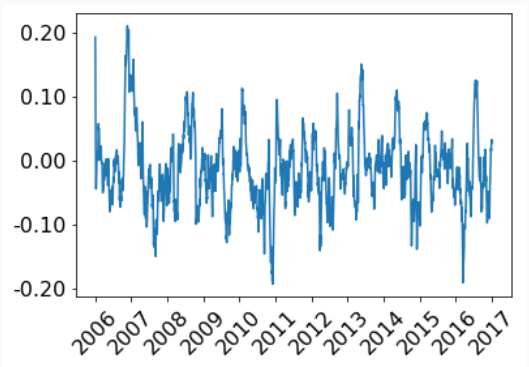
CNN+Transformer Model Structure for Representative Residual Over Time



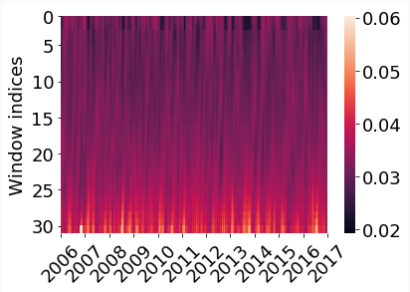
(a) Cumulative residuals



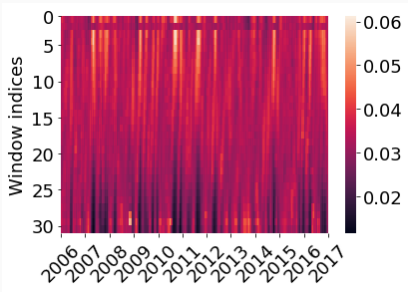
(b) Average attention weights



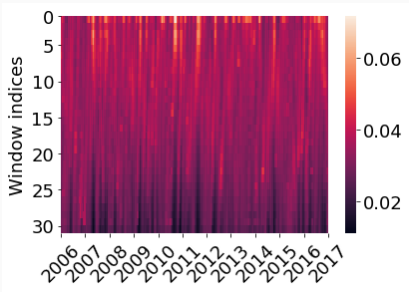
(c) Allocation weights



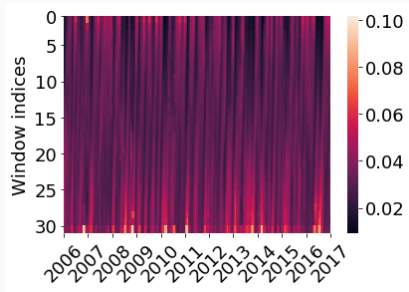
(d) Attention weights for head 1



(e) Attention weights for head 2



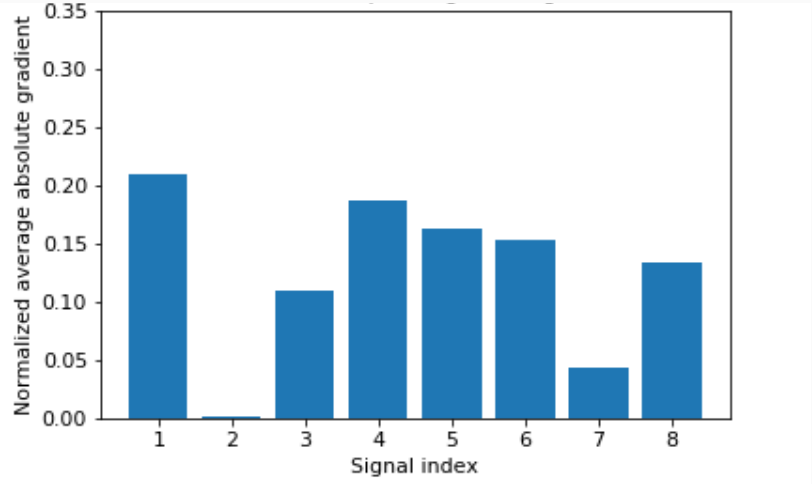
(f) Attention weights for head 3



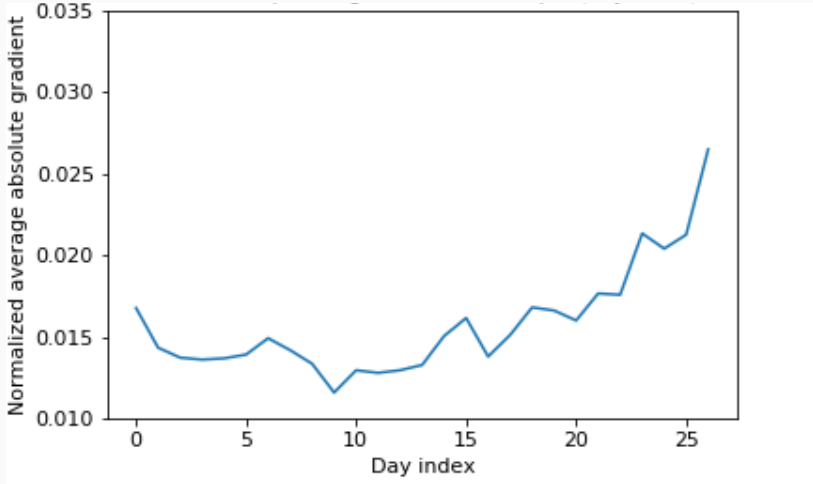
(g) Attention weights for head 4

- Attention head weights 4 highest for down-times in 2009, 2014, middle 2016. Focuses uniformly on last 10 days in 30-day window
- Attention head weights 3 highest for up-patterns in 2007, 2010, 2012. Focuses uniformly on first 20 days in 30-day window
- Asymmetric response of Transformer: act swiftly during downtrends, stay cautious during uptrends

Variable Importance for Allocation Weight



(a) Importance of Local Basic Patterns



(b) Importance of Residual Days

- Measure importance with average absolute gradient of allocation weight
- Most important basic patterns are trends or local curvature.
  Flat basic pattern 2 is negligible.
- All previous days matter, but on average the most recent 14 days get more attention for trading decisions.

Conclusion

Conclusion

Methodology:

- Unifying conceptual framework to compare different approaches:
  (1) portfolio generation, (2) signal extraction, (3) allocation decision
- Novel deep learning statistical arbitrage:
  1. Conditional latent factors to generate arbitrage portfolios
  2. CNN+Transformer signal: global dependency pattern with local filters
  3. FFN allocation and global trading objective for estimation

Empirical results:

- Comprehensive out-of-sample study on U.S. equities
- CNN+Transformer substantially outperforms benchmark approaches
- Unspanned by conventional risk factors
- Survives realistic transaction and holding costs
- Insights into trading policies: asymmetric trend and reversion patterns
- Trading signal extraction is the most challenging and separating element

Appendix

Firm specific characteristics

Past Returns	Investment	Profitability	Intangibles	Value	Trading Frictions
Momentum	Investment	Operating profitability	Accrual	Book to Market Ratio	Size
Short-term Reversal	Net operating assets	Profitability	Operating accruals	Assets to market cap	Turnover
Long-term Reversal	Change in prop. to assets	Sales over assets	Operating leverage	Cash to assets	Idiosyncratic Volatility
Return 2-1	Net Share Issues	Capital turnover	Price to cost margin	Cash flow to book value	CAPM Beta
Return 12-2		Fixed costs to sales		Cashflow to price	Residual Variance
Return 36-13		Profit margin		Dividend to price	Total assets
		Return on net assets		Earnings to price	Market Beta
		Return on assets		Tobin's Q	Close to High
		Return on equity		Sales to price	Spread
		Expenses to sales		Leverage	Unexplained Volume
		Capital intensity			Variance

46 firm-specific monthly characteristics sorted into six categories.

Significance of Arbitrage Alphas

CNN+Trans model

	Fama-French					PCA					IPCA				
K	α	t_α	R^2	μ	t_μ	α	t_α	R^2	μ	t_μ	α	t_α	R^2	μ	t_μ
0	11.6%	6.4***	30.3%	13.7%	6.3***	11.6%	6.4***	30.3%	13.7%	6.3***	11.6%	6.4***	30.3%	13.7%	6.3***
1	7.0%	14***	2.4%	7.2%	14***	14.9%	10***	0.6%	15.2%	11***	8.1%	12***	9.5%	8.7%	12***
3	5.5%	12***	1.2%	5.5%	12***	15.8%	14***	1.7%	16.0%	14***	8.2%	15***	6.0%	8.6%	15***
5	4.5%	12***	2.3%	4.6%	12***	14.1%	13***	1.3%	14.3%	13***	8.3%	16***	3.9%	8.7%	16***
8	3.3%	9.4***	2.1%	3.4%	9.6***	12.0%	12***	0.9%	12.2%	12***	7.8%	15***	5.0%	8.2%	15***
10	-	-	-	-	-	10.5%	11***	0.7%	10.7%	11***	7.7%	15***	4.0%	8.0%	15***
15	-	-	-	-	-	7.5%	8.8***	0.5%	7.6%	8.9***	8.1%	16***	4.2%	8.4%	16***

Fourier+FFN model

	Fama-French					PCA					IPCA				
K	α	t_α	R^2	μ	t_μ	α	t_α	R^2	μ	t_μ	α	t_α	R^2	μ	t_μ
0	2.7%	0.8	8.6%	4.9%	1.4	2.7%	0.8	8.6%	4.9%	1.4	2.7%	0.8	8.6%	4.9%	1.4
1	3.0%	3.3**	3.3%	3.2%	3.5***	7.4%	2.7**	3.3%	8.4%	3.1**	4.8%	4.0***	16.4%	6.3%	4.8***
3	3.2%	4.7***	4.2%	3.5%	5.1***	10.9%	6.3***	2.2%	11.2%	6.4***	6.8%	6.4***	13.0%	7.8%	6.9***
5	2.9%	6.1***	3.5%	3.1%	6.4***	12.1%	7.5***	1.5%	12.4%	7.6***	6.7%	6.9***	13.3%	7.7%	7.4***
8	3.0%	7.2***	3.2%	3.1%	7.4***	10.0%	7.5***	0.9%	10.1%	7.6***	6.8%	7.0***	13.3%	7.8%	7.5***
10	-	-	-	-	-	8.0%	6.5***	1.0%	8.2%	6.6***	6.8%	7.1***	12.7%	7.6%	7.5***
15	-	-	-	-	-	4.7%	4.3***	0.4%	4.8%	4.4***	7.1%	7.6***	12.2%	7.9%	8.0***

OU+Thresh model

	Fama-French					PCA					IPCA				
K	α	t_α	R^2	μ	t_μ	α	t_α	R^2	μ	t_μ	α	t_α	R^2	μ	t_μ
0	-4.5%	-1.4	13.4%	-2.4%	-0.7	-4.5%	-1.4	13.4%	-2.4%	-0.7	-4.5%	-1.4	13.4%	-2.4%	-0.7
1	-0.2%	-0.2	13.5%	0.6%	0.6	0.7%	0.3	6.3%	2.1%	0.8	1.7%	1.4	18.9%	3.0%	2.3*
3	0.9%	1.2	10.4%	1.6%	2.1*	4.3%	2.5*	4.3%	5.2%	3.0**	2.6%	2.6**	18.8%	3.8%	3.4***
5	0.5%	0.9	6.8%	0.9%	1.5	3.7%	2.4*	3.2%	4.4%	2.8**	2.8%	3.0**	17.7%	3.8%	3.8***
8	0.6%	1.2	5.5%	1.0%	1.9	3.9%	3.0**	1.9%	4.4%	3.4***	2.3%	2.6**	17.6%	3.5%	3.6***
10	-	-	-	-	-	2.6%	2.2*	1.4%	2.9%	2.4*	2.1%	2.5*	17.6%	3.1%	3.3***
15	-	-	-	-	-	2.1%	2.1*	0.7%	2.4%	2.4*	2.3%	2.8**	18.1%	3.2%	3.6***

Significance of Arbitrage Alphas with Mean-Variance Objective

CNN+Trans model

	Fama-French					PCA					IPCA				
K	α	t_α	R^2	μ	t_μ	α	t_α	R^2	μ	t_μ	α	t_α	R^2	μ	t_μ
0	5.8%	2.2*	19.6%	9.5%	3.2**	5.8%	2.2*	19.6%	9.5%	3.2**	5.8%	2.2*	19.6%	9.5%	3.2**
1	9.9%	12***	7.1%	10.5%	12***	26.3%	8.3***	1.6%	27.3%	8.6***	14.0%	11***	23.5%	15.9%	11***
3	7.5%	11***	5.3%	7.8%	11***	22.1%	9.1***	2.2%	22.6%	9.2***	16.6%	12***	17.6%	17.9%	12***
5	5.7%	11***	5.3%	5.9%	12***	19.0%	10***	3.2%	19.6%	11***	16.7%	12***	16.0%	18.2%	12***
8	4.4%	9.8***	3.6%	4.6%	10***	16.3%	10***	1.6%	16.6%	10***	15.5%	12***	18.3%	17.0%	12***
10	-	-	-	-	-	14.8%	10***	1.7%	15.3%	10***	15.2%	13***	20.6%	16.6%	12***
15	-	-	-	-	-	8.5%	8.4***	0.9%	8.7%	8.5***	14.8%	13***	21.6%	16.3%	13***

Fourier+FFN model

	Fama-French					PCA					IPCA				
K	α	t_α	R^2	μ	t_μ	α	t_α	R^2	μ	t_μ	α	t_α	R^2	μ	t_μ
0	3.2%	0.7	8.4%	5.5%	1.1	3.2%	0.7	8.4%	5.5%	1.1	3.2%	0.7	8.4%	5.5%	1.1
1	2.8%	1.6	1.8%	2.5%	1.5	15.4%	1.7	1.3%	16.6%	1.9	7.9%	1.8	2.6%	9.7%	2.2*
3	4.1%	4.4***	3.4%	4.3%	4.5***	30.3%	1.3	0.1%	32.1%	1.3	17.4%	4.1***	1.9%	17.6%	4.1***
5	2.9%	4.8***	3.1%	3.1%	5.0***	21.0%	1.3	0.1%	22.5%	1.4	15.9%	4.3***	2.6%	17.0%	4.5***
8	3.5%	6.8***	2.3%	3.6%	7.0***	17.4%	2.6**	0.3%	17.2%	2.6**	12.9%	4.3***	4.4%	14.4%	4.7***
10	-	-	-	-	-	7.1%	1.7	0.3%	7.4%	1.8	11.7%	3.9***	3.5%	12.6%	4.1***
15	-	-	-	-	-	5.5%	2.1*	0.1%	5.7%	2.2*	11.3%	4.3***	4.0%	12.1%	4.5***

Dependency between Arbitrage Strategies

Table 1: Correlations between the Returns of the CNN+Transformer Arbitrage Strategies

	FF 3	PCA 3	IPCA 3	FF 5	PCA 5	IPCA 5	PCA 10	IPCA 10
FF 3	1.00	0.32	0.44	0.62	0.25	0.43	0.21	0.44
PCA 3	0.32	1.00	0.32	0.34	0.62	0.35	0.41	0.36
IPCA 3	0.44	0.32	1.00	0.37	0.28	0.81	0.21	0.75
FF 5	0.62	0.34	0.37	1.00	0.28	0.39	0.23	0.40
PCA 5	0.25	0.62	0.28	0.28	1.00	0.29	0.47	0.31
IPCA 5	0.43	0.35	0.81	0.39	0.29	1.00	0.23	0.84
PCA 10	0.21	0.41	0.21	0.23	0.47	0.23	1.00	0.25
IPCA 10	0.44	0.36	0.75	0.40	0.31	0.84	0.25	1.00

Strategies from different factor models have low inter-family correlations

- Inter-family correlations range from 0.21 to 0.44.
- Intra-family correlations range between 0.41 and 0.84.

Importance of Time-Series Signal

	Factors	Fama-French			PCA			IPCA		
Model	K	SR	μ	σ	SR	μ	σ	SR	μ	σ
OU + FFN	0	0.50	10.6%	21.3%	0.50	10.6%	21.3%	0.50	10.6%	21.3%
	1	0.34	0.8%	2.3%	0.05	0.7%	11.9%	0.60	4.8%	8.0%
	3	0.16	0.2%	1.4%	0.44	3.4%	7.8%	0.70	4.6%	6.6%
	5	0.17	0.2%	1.2%	0.68	4.7%	7.0%	0.66	4.2%	6.3%
	8	-0.34	-0.3%	1.0%	0.31	2.3%	6.9%	0.61	3.9%	6.2%
	10	-	-	-	0.26	1.3%	5.0%	0.56	3.5%	6.2%
	15	-	-	-	0.31	1.4%	4.3%	0.54	3.3%	6.1%
FFN	0	0.57	8.8%	15.3%	0.57	8.8%	15.3%	0.57	8.8%	15.3%
	1	0.60	2.0%	3.3%	0.53	6.2%	11.7%	1.07	6.5%	6.1%
	3	1.02	2.6%	2.6%	1.15	8.2%	7.2%	1.50	7.6%	5.0%
	5	1.32	2.3%	1.7%	1.42	9.8%	6.9%	1.55	7.3%	4.7%
	8	1.31	2.1%	1.6%	0.84	5.1%	6.1%	1.56	7.2%	4.6%
	10	-	-	-	0.70	3.5%	5.0%	1.48	7.0%	4.7%
	15	-	-	-	0.51	2.4%	4.8%	1.68	7.5%	4.5%

Robustness to Rolling Window Size

Table 2: OOS Annualized Performance of CNN+Trans for 60 Days Lookback Window

	Fama-French			PCA			IPCA		
K	SR	μ	σ	SR	μ	σ	SR	μ	σ
0	1.50	13.5%	9.0%	1.50	13.5%	9.0%	1.50	13.5%	9.0%
1	2.95	9.6%	3.2%	2.68	15.8%	5.9%	3.14	8.8%	2.8%
3	3.21	8.7%	2.7%	3.49	16.8%	4.8%	3.84	9.6%	2.5%
5	3.23	6.8%	2.1%	3.54	16.0%	4.5%	3.90	9.2%	2.4%
8	2.96	4.2%	1.4%	3.02	12.5%	4.2%	3.93	8.7%	2.2%
10	-	-	-	2.67	9.9%	3.7%	3.98	9.2%	2.3%
15	-	-	-	2.36	8.1%	3.4%	4.24	9.6%	2.3%

Robustness to Rolling Window Size

Table 3: Significance of Arbitrage Alphas for 60 Days Lookback Window

CNN+Trans Model , Sharpe objective function, $L = 60$ days lookback window															
	Fama-French					PCA					IPCA				
K	α	t_α	R^2	μ	t_μ	α	t_α	R^2	μ	t_μ	α	t_α	R^2	μ	t_μ
0	11.8%	5.6***	19.5%	13.5%	5.8***	11.8%	5.6***	19.5%	13.5%	5.8***	11.8%	5.6***	19.5%	13.5%	5.8***
1	9.1%	11***	7.2%	9.6%	11***	15.5%	10***	1.2%	15.8%	10***	8.2%	12***	10.1%	8.8%	12***
3	8.3%	12***	7.1%	8.7%	12***	16.5%	13***	2.5%	16.8%	14***	9.2%	15***	9.3%	9.6%	15***
5	6.5%	12***	6.0%	6.8%	13***	15.6%	13***	2.2%	16.0%	14***	8.8%	15***	10.3%	9.2%	15***
8	4.1%	11***	3.2%	4.2%	11***	12.2%	11***	1.6%	12.5%	12***	8.3%	15***	8.9%	8.7%	15***
10	-	-	-	-	-	9.7%	10***	1.0%	9.9%	10***	8.8%	15***	8.3%	9.2%	15***
15	-	-	-	-	-	8.1%	9.1***	0.7%	8.1%	9.1***	9.2%	16***	9.3%	9.6%	16***

Mean-Variance Objective

CNN+Trans model, mean-variance objective function

	Fama-French			PCA			IPCA		
K	SR	μ	σ	SR	μ	σ	SR	μ	σ
0	0.83	9.5%	11.4%	0.83	9.5%	11.4%	0.83	9.5%	11.4%
1	3.15	10.5%	3.3%	2.21	27.3%	12.3%	2.83	15.9%	5.6%
3	2.95	7.8%	2.6%	2.38	22.6%	9.5%	3.13	17.9%	5.7%
5	3.03	5.9%	2.0%	2.75	19.6%	7.1%	3.21	18.2%	5.7%
8	2.96	4.2%	1.4%	2.68	16.6%	6.2%	3.18	17.0%	5.4%
10	-	-	-	2.67	15.3%	5.7%	3.21	16.6%	5.2%
15	-	-	-	2.20	8.7%	4.0%	3.34	16.3%	4.9%

Fourier+FFN model, mean-variance objective function

	Fama-French			PCA			IPCA		
K	SR	μ	σ	SR	μ	σ	SR	μ	σ
0	0.28	5.5%	19.3%	0.28	5.5%	19.3%	0.28	5.5%	19.3%
1	0.38	2.5%	6.7%	0.48	16.6%	34.8%	0.56	9.7%	17.2%
3	1.16	4.3%	3.7%	0.34	32.1%	93.1%	1.06	17.6%	16.7%
5	1.30	3.1%	2.4%	0.37	22.5%	61.2%	1.17	17.0%	14.5%
8	1.73	3.6%	2.0%	0.67	17.4%	25.9%	1.21	14.4%	11.9%
10	-	-	-	0.45	7.4%	16.4%	1.06	12.6%	11.9%
15	-	-	-	0.56	5.7%	10.2%	1.17	12.1%	10.4%

Constant Model without Re-estimation

Table 4: OOS Annualized Performance of CNN+Trans for Constant Model

$T_{\text{train}} = 4$ years									
	Fama-French			PCA			IPCA		
K	SR	μ	σ	SR	μ	σ	SR	μ	σ
0	1.10	8.5%	7.8%	1.10	8.5%	7.8%	1.10	8.5%	7.8%
1	1.90	4.5%	2.3%	0.44	3.0%	6.9%	0.94	3.1%	3.3%
3	1.60	3.6%	2.2%	1.65	8.7%	5.3%	1.82	5.3%	2.9%
5	1.81	3.0%	1.7%	1.93	9.8%	5.1%	2.09	5.4%	2.6%
8	1.70	2.5%	1.5%	2.04	9.6%	4.7%	1.89	5.0%	2.6%
10	-	-	-	2.06	9.1%	4.4%	1.77	4.7%	2.7%
15	-	-	-	1.82	7.0%	3.9%	2.09	5.5%	2.7%
$T_{\text{train}} = 8$ years									
	Fama-French			PCA			IPCA		
K	SR	μ	σ	SR	μ	σ	SR	μ	σ
0	1.33	12.0%	9.0%	1.33	12.0%	9.0%	1.33	12.0%	9.0%
1	2.06	5.0%	2.4%	1.81	15.2%	8.4%	2.02	8.5%	4.2%
3	2.46	5.3%	2.2%	2.04	13.1%	6.4%	2.47	7.5%	3.0%
5	1.82	3.2%	1.8%	1.91	11.9%	6.2%	2.64	7.6%	2.9%
8	1.48	2.5%	1.7%	1.89	10.8%	5.7%	2.71	8.3%	3.1%
10	-	-	-	1.82	10.0%	5.5%	2.68	8.2%	3.1%
15	-	-	-	1.38	6.2%	4.5%	2.70	7.8%	2.9%

Constant Model without Re-estimation

Table 5: Significance of Arbitrage Alphas for Constant Model

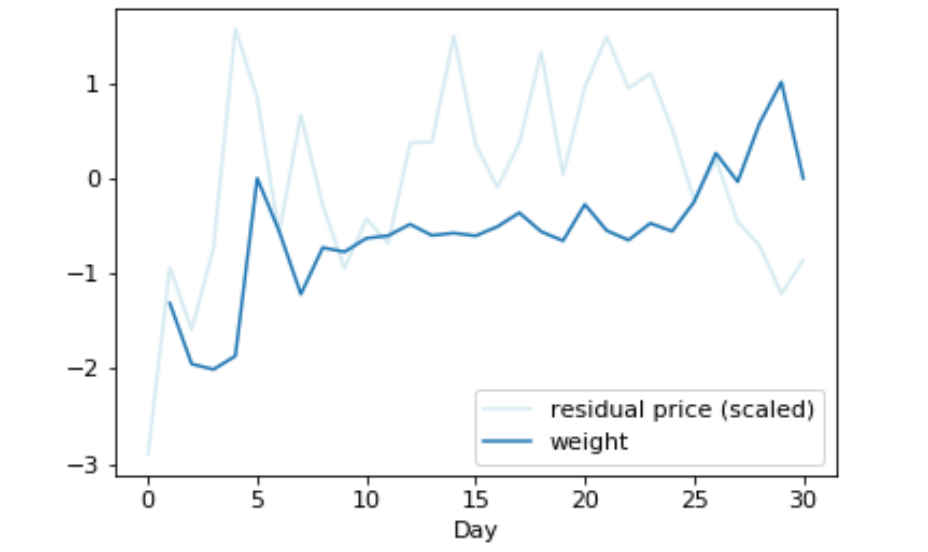
CNN+Trans model, Sharpe objective function, $T_{\text{train}} = 4$ years

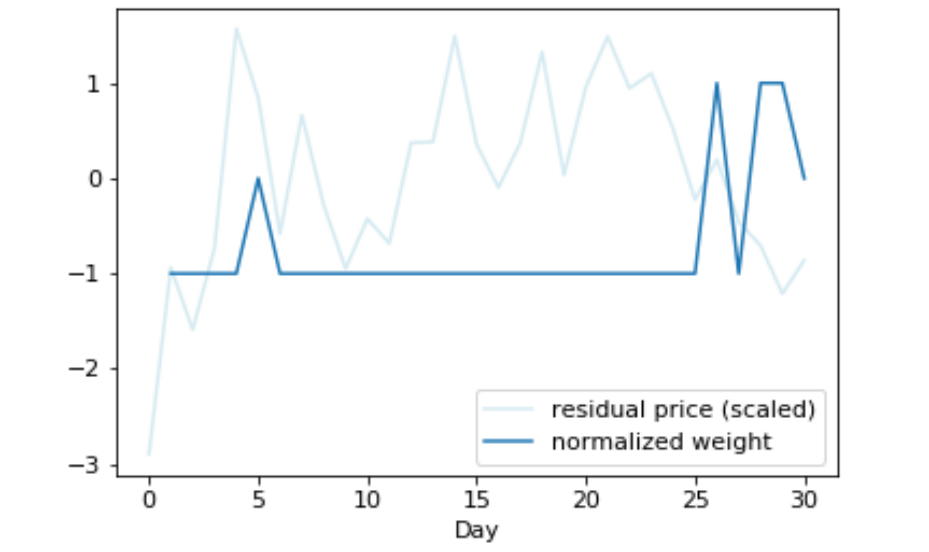
	Fama-French					PCA					IPCA				
K	α	t_α	R^2	μ	t_μ	α	t_α	R^2	μ	t_μ	α	t_α	R^2	μ	t_μ
0	8.4%	4.2^{***}	3.0%	8.5%	4.3^{***}	8.4%	4.2^{***}	3.0%	8.5%	4.3^{***}	8.4%	4.2^{***}	3.0%	8.5%	4.3^{***}
1	4.0%	6.8^{***}	5.9%	4.5%	7.3^{***}	4.1%	2.0^{*}	4.5%	5.2%	2.5^{*}	3.1%	3.7^{***}	1.6%	3.1%	3.6^{***}
3	3.2%	5.7^{***}	4.9%	3.6%	6.2^{***}	8.2%	6.1^{***}	2.7%	8.7%	6.4^{***}	5.3%	7.4^{***}	11.7%	5.3%	7.0^{***}
5	2.8%	6.6^{***}	4.3%	3.0%	7.0^{***}	9.3%	7.1^{***}	1.8%	9.8%	7.5^{***}	5.5%	8.6^{***}	8.3%	5.4%	8.1^{***}
8	2.3%	6.1^{***}	5.1%	2.5%	6.6^{***}	9.0%	7.5^{***}	2.2%	9.6%	7.9^{***}	5.0%	7.7^{***}	8.2%	5.0%	7.3^{***}
10	-	-	-	-	-	8.6%	7.5^{***}	1.9%	9.1%	8.0^{***}	5.1%	8.0^{***}	16.6%	4.7%	6.9^{***}
15	-	-	-	-	-	6.8%	6.8^{***}	1.0%	7.0%	7.1^{***}	5.8%	9.3^{***}	17.6%	5.5%	8.1^{***}

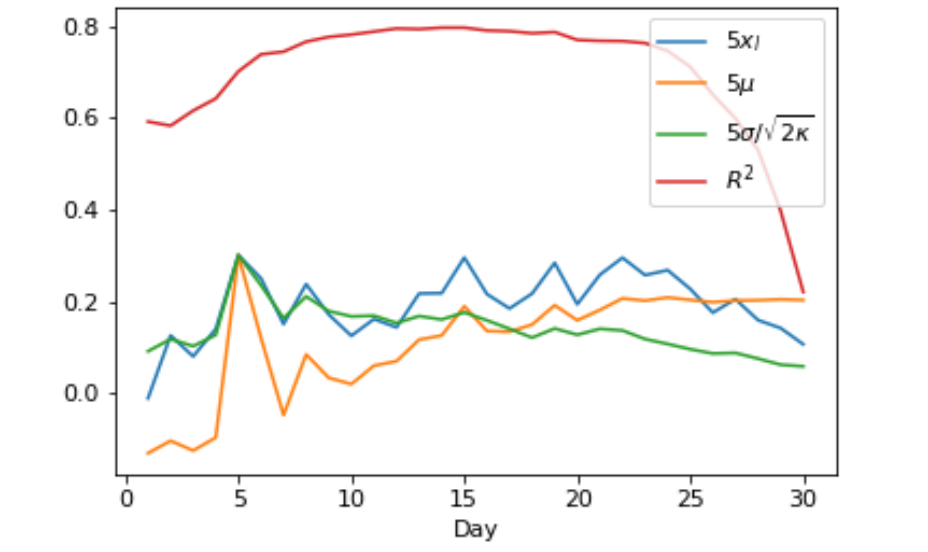
CNN+Trans model, Sharpe objective function, $T_{\text{train}} = 8$ years

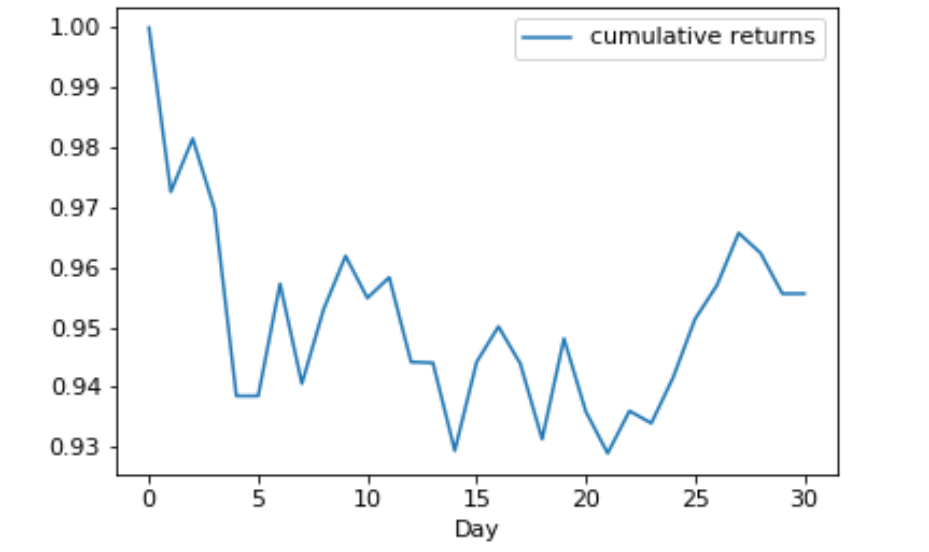
	Fama-French					PCA					IPCA				
K	α	t_α	R^2	μ	t_μ	α	t_α	R^2	μ	t_μ	α	t_α	R^2	μ	t_μ
0	10.1%	4.1^{***}	18.1%	12.0%	4.4^{***}	10.1%	4.1^{***}	18.1%	12.0%	4.4^{***}	10.1%	4.1^{***}	18.1%	12.0%	4.4^{***}
1	4.4%	6.5^{***}	14.3%	5.0%	6.8^{***}	14.5%	5.8^{***}	2.5%	15.2%	6.0^{***}	7.0%	6.6^{***}	30.6%	8.5%	6.7^{***}
3	4.9%	7.9^{***}	11.6%	5.3%	8.2^{***}	12.8%	6.7^{***}	2.7%	13.1%	6.8^{***}	7.0%	7.9^{***}	8.2%	7.5%	8.2^{***}
5	2.9%	5.8^{***}	12.3%	3.2%	6.0^{***}	11.6%	6.2^{***}	1.6%	11.9%	6.3^{***}	7.1%	8.7^{***}	12.1%	7.6%	8.7^{***}
8	2.3%	4.7^{***}	5.4%	2.5%	4.9^{***}	10.2%	6.0^{***}	3.1%	10.8%	6.3^{***}	7.7%	9.0^{***}	14.6%	8.3%	9.0^{***}
10	-	-	-	-	-	9.4%	5.7^{***}	2.6%	10.0%	6.0^{***}	7.7%	8.9^{***}	11.3%	8.2%	8.9^{***}
15	-	-	-	-	-	6.0%	4.4^{***}	0.9%	6.2%	4.6^{***}	7.4%	8.9^{***}	11.2%	7.8%	8.9^{***}

Empirical example: (1) OU+Threshold signals & allocation weights

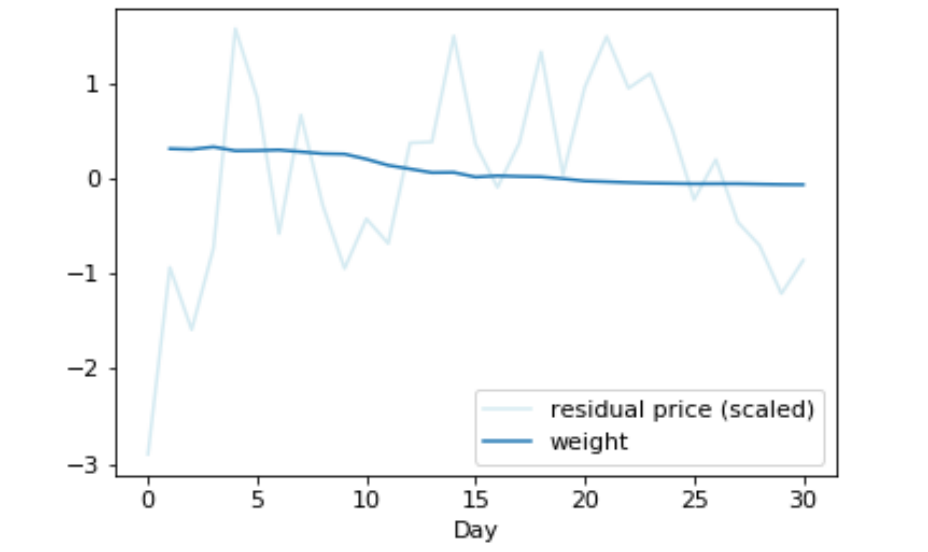


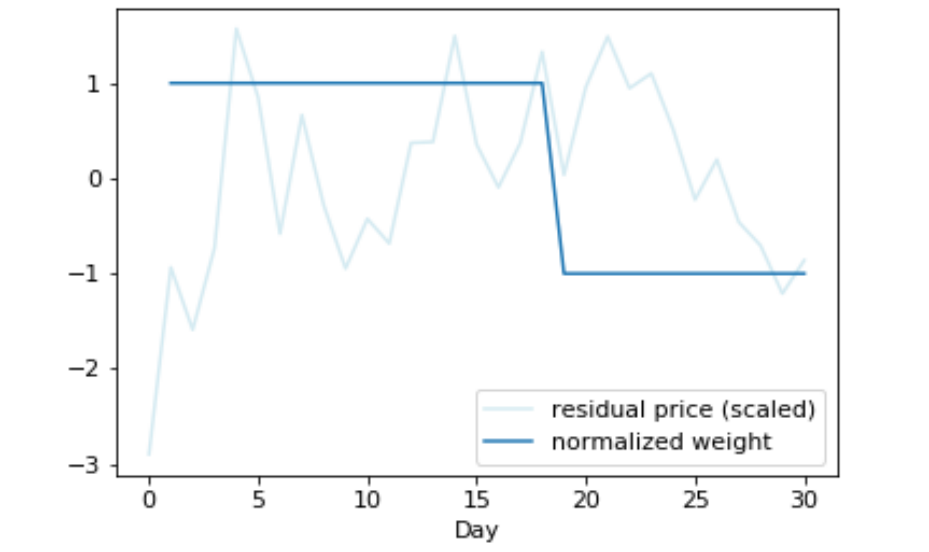


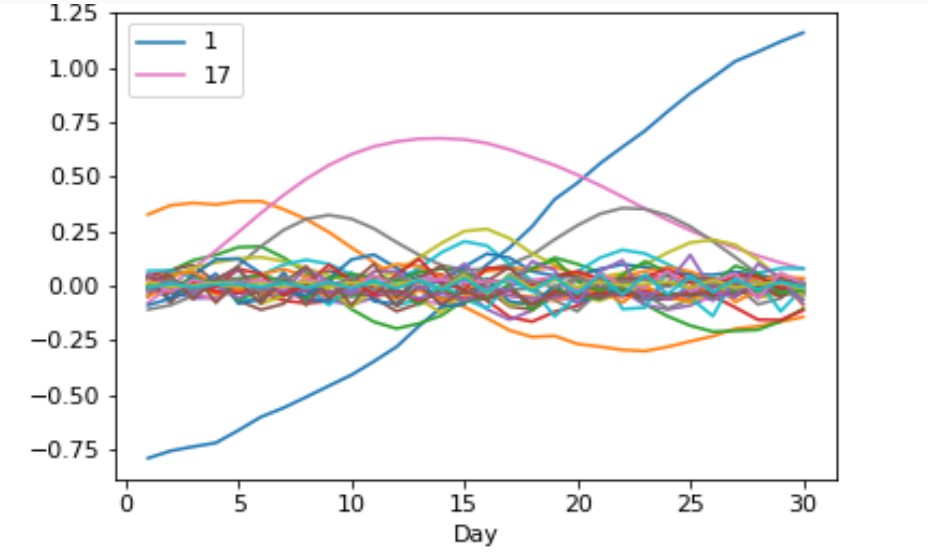


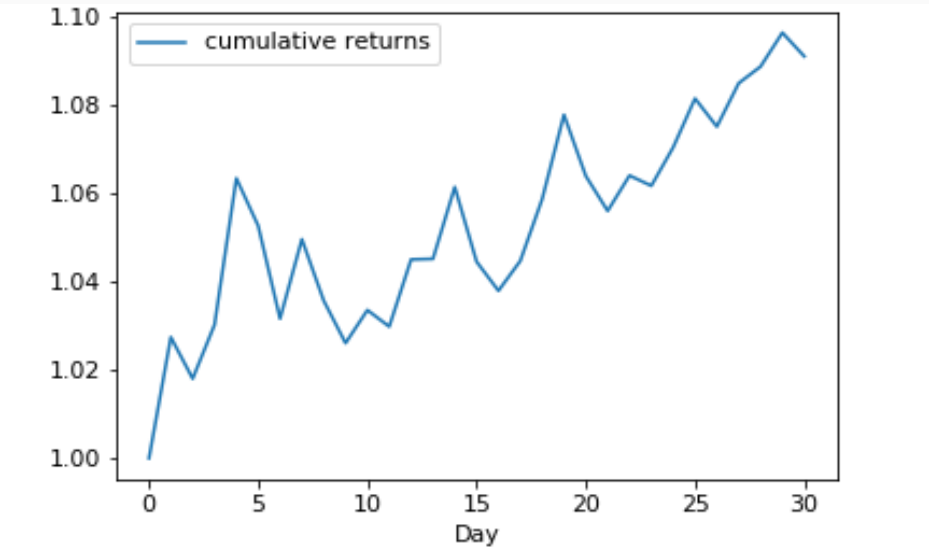


Empirical example: (2) Fourier+FFN signals & allocation weights

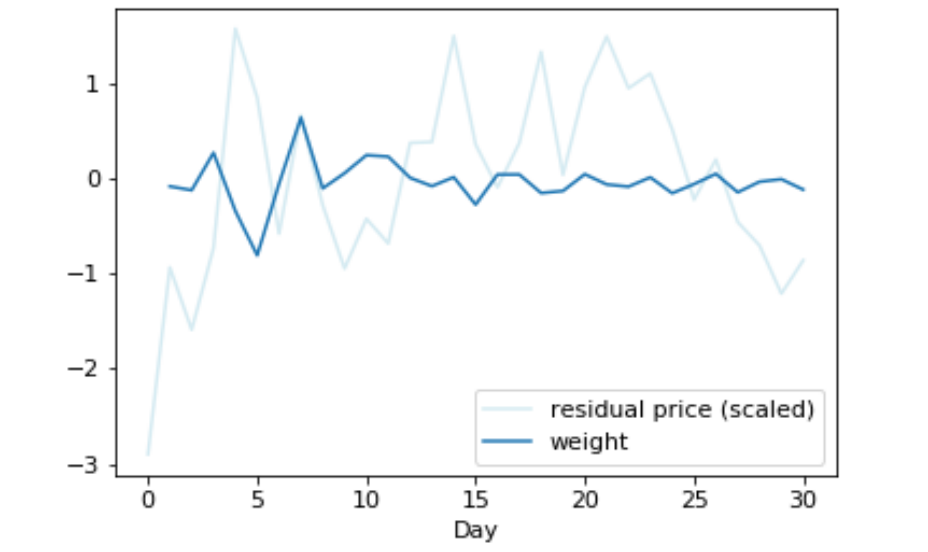


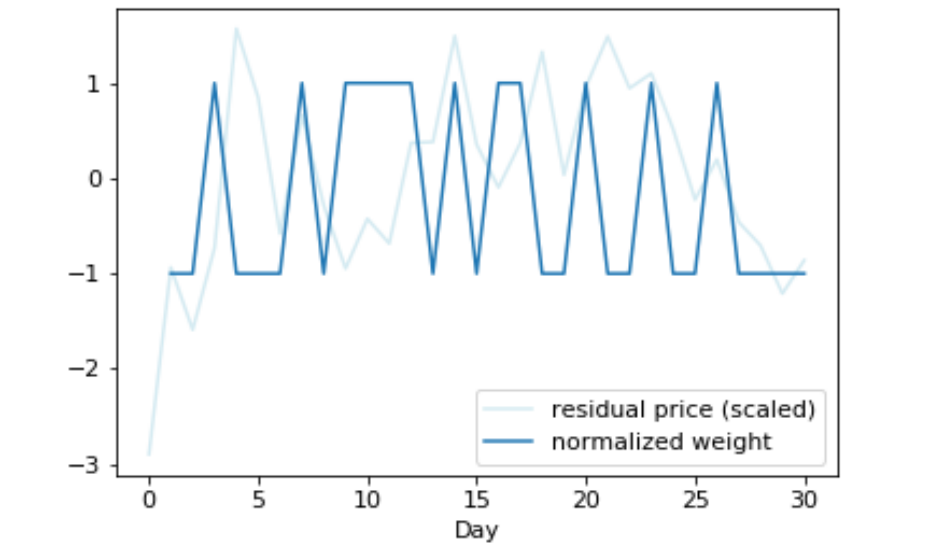


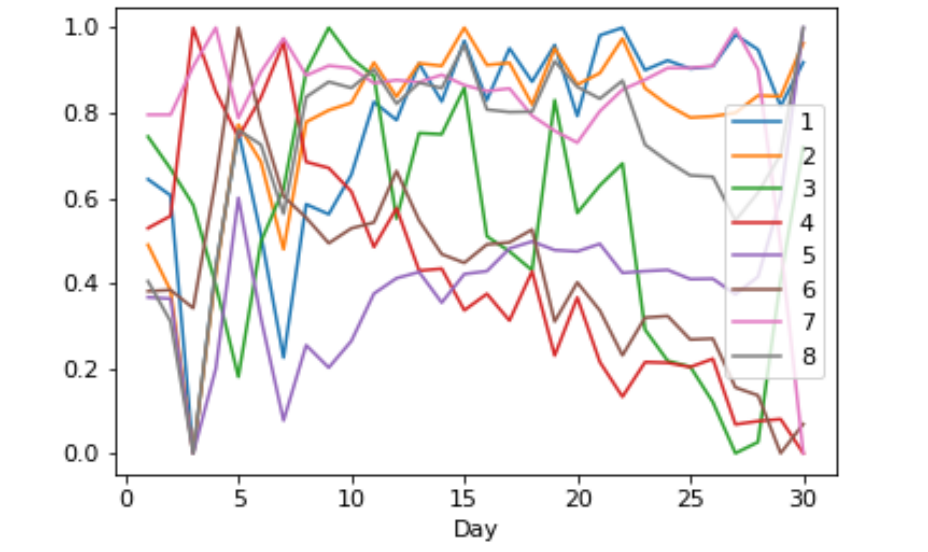


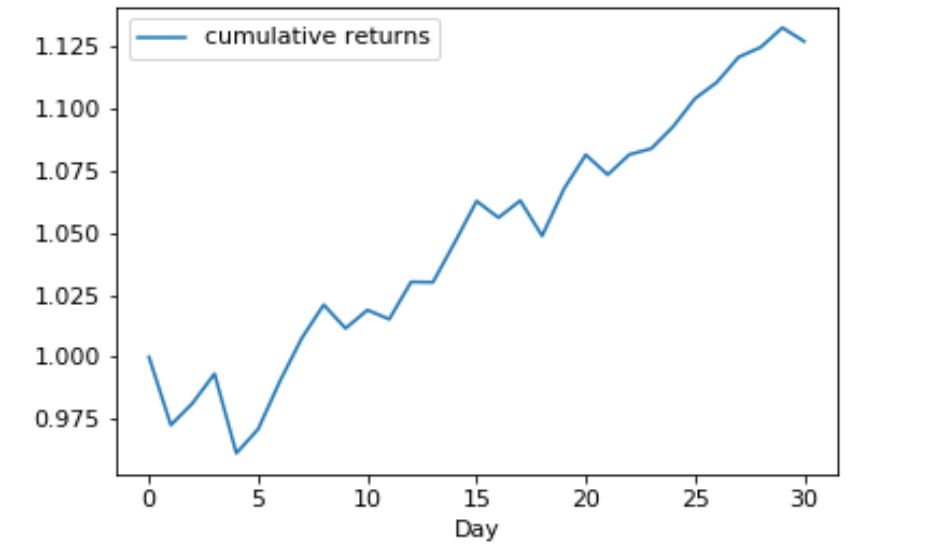


Simulation example: (3) CNN+Transformer signals & allocation weights

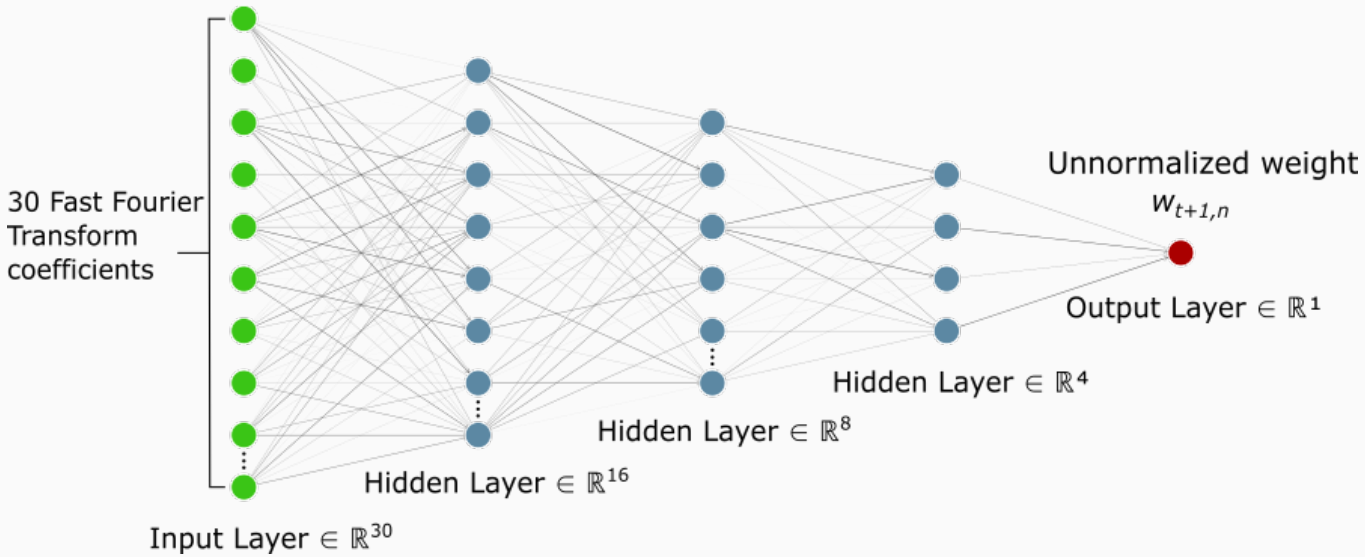








Fourier+FFN architecture



FFN equations:

$$x^{(l)} = \text{ReLU}(W^{(l-1)}x^{(l-1)} + b^{(l-1)})$$

$$w = W^{(L)}x^{(L)} + b^{(L)}$$

Convolutional network equations

Given

$$\mu_k^{(i)} = \frac{1}{L}\sum_{l=1}^{L} y_{l,k}^{(i)}, \quad \sigma_k^{(i)} = \sqrt{\frac{1}{L}\sum_{l=1}^{L}\left(y_{l,k}^{(i)} - \mu_k^{(i)}\right)^2}.$$

Input time series $x^{(0)} \in \mathbb{R}^L$ are passed through $k = 1, \ldots, F$ convolutional filters of size f_{size}, normalization, and ReLU:

$$y_{l,k}^{(0)} = b_k^{(1)} + \sum_{m=1}^{f_{size}} W_{k,m}^{(1)} x_{l-m+1}^{(0)}, \quad x_{l,k}^{(1)} = \mathrm{ReLU}\left(\frac{y_{1,k}^{(0)} - \mu_k^{(0)}}{\sigma_k^{(0)}}\right).$$

Output $x_1^{(1)} \in \mathbb{R}^{L \times F}$ passes through $k = 1, \ldots, F$ convolutional filters of size $f_{size} \times F$, normalization, and ReLU:

$$y_{l,k}^{(1)} = b_k^{(2)} + \sum_{m=1}^{f_{size}} \sum_{k=1}^{F} W_{k,m}^{(2)} x_{l-m+1,k}^{(1)}, \quad x_{l,k}^{(2)} = \mathrm{ReLU}\left(\frac{y_{l,k}^{(1)} - \mu_k^{(1)}}{\sigma_k^{(1)}}\right),$$

Finally, residuals are added back to output $x^{(2)} \in \mathbb{R}^{L \times F}$ via residual connection, to compute features $\tilde{x} \in \mathbb{R}^{L \times F}$:

$$\tilde{x}_{l,k} = x_{l,k}^{(2)} + x_l^{(0)}.$$

$\Rightarrow$ All $b^{(i)}$ and $W^{(i)}$ are parameters; all convolutions are left-padded with 0.

Transformer equations

- Features $\tilde{x} \in \mathbb{R}^{L \times F}$ are projected onto $i = 1, \ldots, h$ F/h-dimensional subspaces ("heads"):

$$V_i = \tilde{x} W_i^V + b_i^V, \quad K_i = \tilde{x} W_i^K + b_i^K, \quad Q_i = \tilde{x} W_i^Q + b_i^Q \in \mathbb{R}^{L \times F/h}.$$

- Projections V_i are aggregated temporally obtaining hidden states $y_i \in \mathbb{R}^{L \times F/h}$, for

$$y_{i,l} = \sum_{j=1}^{L} w_{l,j}^{(i)} V_{i,j} \in \mathbb{R}^{F/h}, \quad w_{l,j}^{(i)} = \frac{\exp(K_{i,l} \cdot Q_{i,j})}{\sum_{m=1}^{L} \exp(K_{i,l} \cdot Q_{i,m})} \in [0, 1].$$

- Final output is $\mathrm{Concat}(y_1, ..., y_h) W^O + b^O \in \mathbb{R}^{L \times F}$.
- This is passed through time-wise feedforward networks.
- $W_i^V, W_i^K, W_i^Q \in \mathbb{R}^{F \times F/h}$, $b_i^V, b_i^K, b_i^Q \in \mathbb{R}^{F/h}$, $W^O \in \mathbb{R}^{F \times F}, b^O \in \mathbb{R}^F$ $\Rightarrow$ parameters to estimate.

Hyperparameter information

Notation	Model Hyperparameters	Initial	Candidates	Chosen
	OU+Thres			
R2T	R^2 filter threshold	0.5	0.25, 0.5, 0.75	0.25
ST	Signal threshold to long/short	1.25	1, 1.25, 1.5	1.25
LKB	Number of days in residual lookback window	30	30	30
	DFT+FFN			
HLC	Hidden layer configuration	[16,8,4]	[16,8,4]	[16,8,4]
DRPH	Dropout rate (% removed) in hidden layers	0.25	0.25	0.25
LKB	Number of days in residual lookback window	30	30	30
WDW	Number of days in rolling training window	1000	1000	1000
RTFQ	Number of days of retraining frequency	125	125	125
	CNN+Trans			
D	Number of filter channels in CNN	4	4, 8	8
ATT	Number of attention heads	4	2, 4	4
HDN	Number of hidden units in transformer's linear layer	2F	2F, 3F	2F
DRPA	Dropout rate (% removed) in the transformer	0.25	0.5, 0.25	0.25
D_{size}	Filter size in CNN	2	2	2
LKB	Number of days in residual lookback window	30	30	30
WDW	Number of days in rolling training window	1000	1000	1000
RTFQ	Number of days of retraining frequency	125	125	125