A Guide to Factor-based Investing

Jongho Kim

Cornell University

SC Johnson Graduate School of Management

Factor Investing

- The core aim of factor models is to understand the drivers of asset prices

- Which characteristics really provide independent information about average returns?

The Workflow of Factor Models

- Step1. Universe filtering
- Step2. Sorting based on a particular factor (e.g., size, book-to-market ratio)
- Step3. Quantile portfolio construction (J=2, J=3, J=5 or J=10 portfolios)
- Step4. Weighting schemes (Equally weighted, value weighted, risk parity)
- Step5. Report the returns of the portfolios (T-test)

Step1. Universe Filtering

- (Terminology) Universe
  - Equities composing the benchmark (e.g. S&P500, KOSPI 200)

- Why do we need universe filters?
  - Reduce potential “Luck” components
  - Building a tradable stock universe (Backtesting = Real Investments)
    - Transaction Costs
    - High Turnovers
    - Timing Loss

Step1. Universe Filtering

- Examples of universe filters
  - Market capitalization
  - Trading dollar volume
  - Days from IPO
  - etc

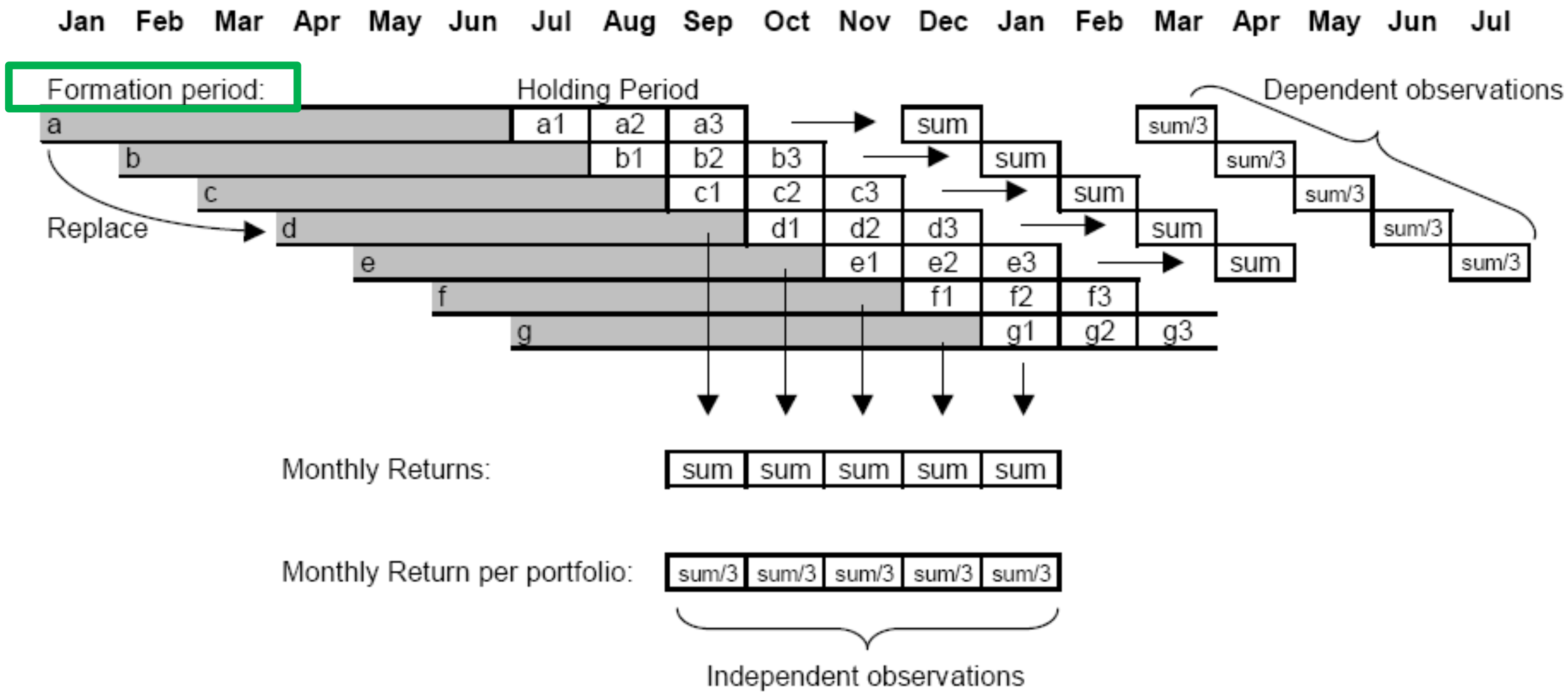
Step2. Sorting Based on a Particular Factor

- Rank firms according to a particular criterion at formation period
  - e.g., momentum, size, book-to-market ratio
  - momentum: sort portfolio based on 12-1M return

https://quant.stackexchange.com/questions/26382/how-to-calculate-monthly-return-from-a-momentum-strategy-with-overlapping-holdin

Sub Portfolio Construction

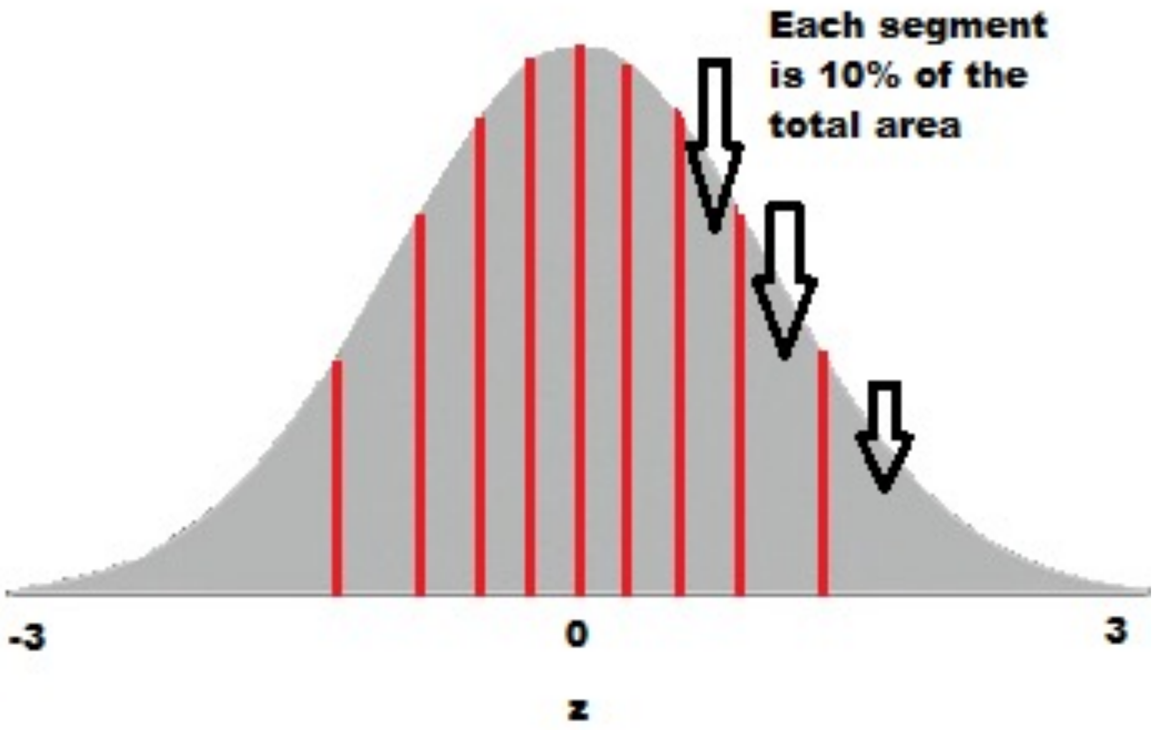
- Formation periods: 1Y
- Holding periods: 1Y
- Rebalancing periods: 1M



Step3. Quantile Portfolio Construction

- Form J ≥ 2 portfolios (i.e., homogeneous groups) consisting of the same number of stocks according to the ranking

  - Ex. Decile



Step4. Weighting Schemes

- Equally weighted

$$W_i = \frac{1}{N}$$

$$N = Total\ number\ of\ securities\ in\ the\ quantile$$
$$W_i = weight\ of\ security$$

  - May have higher transaction costs due to frequent rebalancing required to maintain equal weights
  - Have more weights to small-cap than value weighted

Step4. Weighting Schemes

- Value weighted

$$W_i = \frac{Market\ Cap\ of\ Security\ i}{Total\ Market\ Cap\ of\ the\ Quantile}$$

  - Accounts for market capitalization and weights stocks based on their market value, which means larger companies have a larger impact on the portfolio
  - Can lead to concentration risk, as the portfolio may be heavily weighted towards a few large companies or sectors

Step4. Weighting Schemes

- Risk parity

$$W_i = \frac{\sigma_i^{-1}}{\Sigma\sigma_j^{-1}}$$

Assuming that the assets are uncorrelated

$$\sigma_i = Standard\ deviation\ of\ security\ i$$

  - Designed to provide equal risk exposure across asset classes, which can lead to a more stable portfolio in volatile market conditions
  - Can potentially provide better risk-adjusted returns compared to other weighting schemes
  - Diversifies the portfolio beyond traditional asset classes and can include alternative investments such as commodities or real estate

Step5. Report the Returns of the Portfolios

- The core purpose of factors is to explain the cross-section of stock returns

- An anomaly is identified if the t-test between the first ($j = 1$) and the last group ($j = J$) unveils a significant difference in average returns

Step5. Report the Returns of the Portfolios

Table IV

Momentum Strategy Returns

This table reports the mean quintile portfolio returns based on the past one-week, two-week, three-week, four-week, and one-to-four-week return measures. The mean returns are the time-series averages of weekly value-weighted portfolio excess returns. *, **, and *** denote significance at the 10%, 5%, and 1% levels.

	Quintiles					
	1	2	3	4	5	5–1
r 1,0	**Low**				**High**	
Mean	–0.002	0.000	0.010	0.036**	0.023**	0.025**
t(Mean)	(–0.19)	(0.04)	(1.45)	(2.52)	(2.03)	(2.19)
r 2,0	**Low**				**High**	
Mean	0.000	0.005	0.009	0.017**	0.031***	0.031***
t(Mean)	(0.01)	(0.66)	(1.33)	(2.15)	(2.93)	(2.90)
r 3,0	**Low**				**High**	
Mean	0.005	0.002	0.016*	0.017**	0.036***	0.031***
t(Mean)	(0.60)	(0.28)	(1.94)	(2.30)	(3.21)	(2.65)
r 4,0	**Low**				**High**	
Mean	0.002	0.005	0.009	0.020**	0.025**	0.022**
t(Mean)	(0.30)	(0.66)	(1.28)	(2.45)	(2.32)	(2.26)
r 4,1	**Low**				**High**	
Mean	0.003	0.007	0.021**	0.011	0.020**	0.017*
t(Mean)	(0.35)	(0.94)	(2.34)	(1.51)	(2.02)	(1.82)

Step5. Report the Returns of the Portfolios

- Also, need to evaluate long only position
  - In most cases, portfolios take long only positions, so we cannot exploit (long - short factor) returns
    - Long only position = market factor + factor tilts
  - Benchmark for long only position
    - Market Cap Weighted
    - Equally Weighted
    - Risk Parity

The Workflow of Factor Models

- Step1. Universe filtering
- Step2. Sorting based on a particular factor (e.g., size, book-to-market ratio)
- Step3. Quantile portfolio construction (J=2, J=3, J=5 or J=10 portfolios)
- Step4. Weighting schemes (Equally weighted, value weighted, risk parity)
- Step5. Report the returns of the portfolios (T-test)

Properties of Good Factors

- **Causality is key**
  - If one is able to identify X → expected return, then the problem is solved
  - Unfortunately, causality is incredibly hard to uncover.

- **Financial datasets are extremely noisy**
  - No-arbitrage reasonings imply that if a simple pattern yielded durable profits, it would mechanically and rapidly vanish.

- To maximize out-of-sample accuracy, the right question is: **what’s not going to change?**

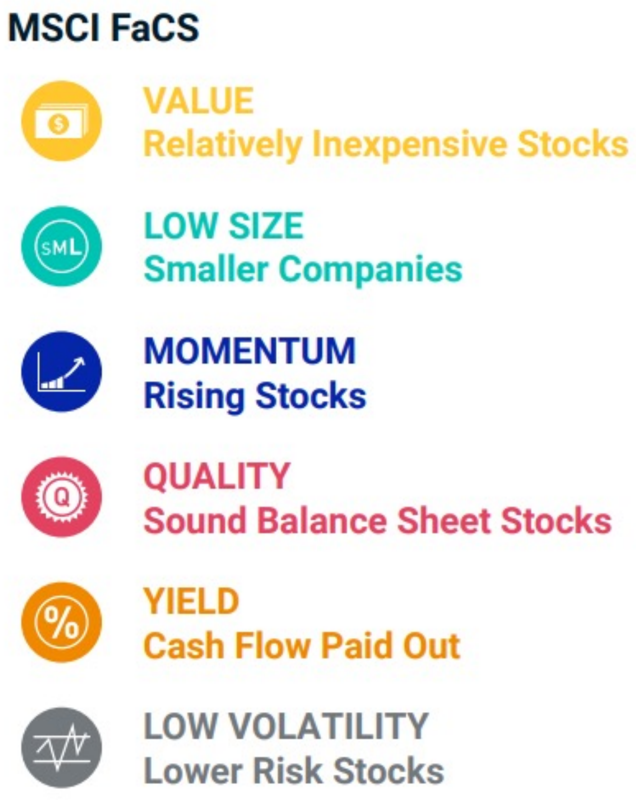
Underlying Mechanisms of Risk Premiums

- **Type1. Risk Compensation**
  - Compensation from exposures to systemic risk

- **Type2. Mispricing**
  - Efficient Market Hypothesis (EMH)
    - Null Hypothesis: News is rapidly and fully incorporated in prices

Examples of Common Factor Groups

- A factor is simply a systematic way of ranking (and selecting) stocks. It could be as simple as value (e.g., P/E) or momentum (e.g., past 12-month returns).



Factor Groups	What it is
Value	Value stocks are those that are considered undervalued compared to their fundamentals, such as earnings or book value. (ex. PBR, PER)
Size	The size factor is based on the observation that smaller companies tend to outperform larger ones over time.
Momentum	Momentum investing involves buying stocks with strong recent performance and selling those with weak performance.
Quality	Quality stocks are characterized by high profitability, low leverage, and stable earnings. (ex. ROE)
Yield	Captures excess returns to stocks that have higher-than-average dividend yields
Low `risk'	This factor is based on the observation that stocks with lower volatility tend to perform better on a risk-adjusted basis. (ex. volatility, market beta, idiosyncratic volatility, etc.)

https://www.msci.com/documents/10199/904e031c-94e4-4dbc-a314-7c373446dffa

Underlying Mechanisms of Risk Premiums

Factor Groups	Underlying Mechanisms
Value	The value premium is often attributed to risk compensation, as value stocks may be more sensitive to economic downturns or mispricing due to investors' behavioral biases, such as overreaction to negative news or short-term earnings fluctuations.
Size	This premium is often attributed to risk compensation, as smaller companies may have higher business risks and lower liquidity. It can also be due to mispricing because of investors' biases, such as neglecting smaller firms in favor of well-known, larger companies.
Momentum	The momentum effect is generally seen as a result of mispricing, driven by investors' behavioral biases, such as underreaction to new information or anchoring on past prices, leading to trends that continue for some time.
Quality	The quality premium is often seen as a result of risk compensation, as higher-quality companies may be more resilient during economic downturns. It could also be due to mispricing, as investors might underestimate the long-term benefits of high-quality businesses.
Yield	The dividend yield premium is typically seen as a result of risk compensation, as high-dividend stocks may be more stable and mature companies. It could also be due to mispricing if investors undervalue the long-term benefits of dividend-paying stocks, such as compounding returns through reinvested dividends.
Low `risk’	The low volatility premium can be attributed to risk compensation because less volatile stocks may offer more downside protection in turbulent markets. It can also be due to mispricing, as investors might prefer high-volatility stocks due to overconfidence or the "lottery ticket" effect, where they chase high returns.

Limitations of Sorted Portfolios

- The sorting criterion could have a non-monotonic (non-linear) impact on returns
- Another concern is that these sorted portfolios may capture not only the priced risk associated to the characteristic, but also some unpriced risk (Control contributions of other factors)

Factor Model: Cross-sectional regression

$X_{t,n,k}$

Exposure of factor k to stock n at time t

Factor Model: Cross-sectional regression

Return of Factor k at time t (Common variables across different n)

Return $\boldsymbol{r_{t,n}} = \Sigma\, \mathrm{x}_{t,n,k} f_{t,k} + \varepsilon_{t,n}$

Exposure of factor k to stock n at time t

Factor Model: Cross-sectional regression

Return of Factor k at time t (Common variables across different n)

Return $r_{t,n} = \Sigma\, x_{t,n,k} f_{t,k} + \varepsilon_{t,n}$

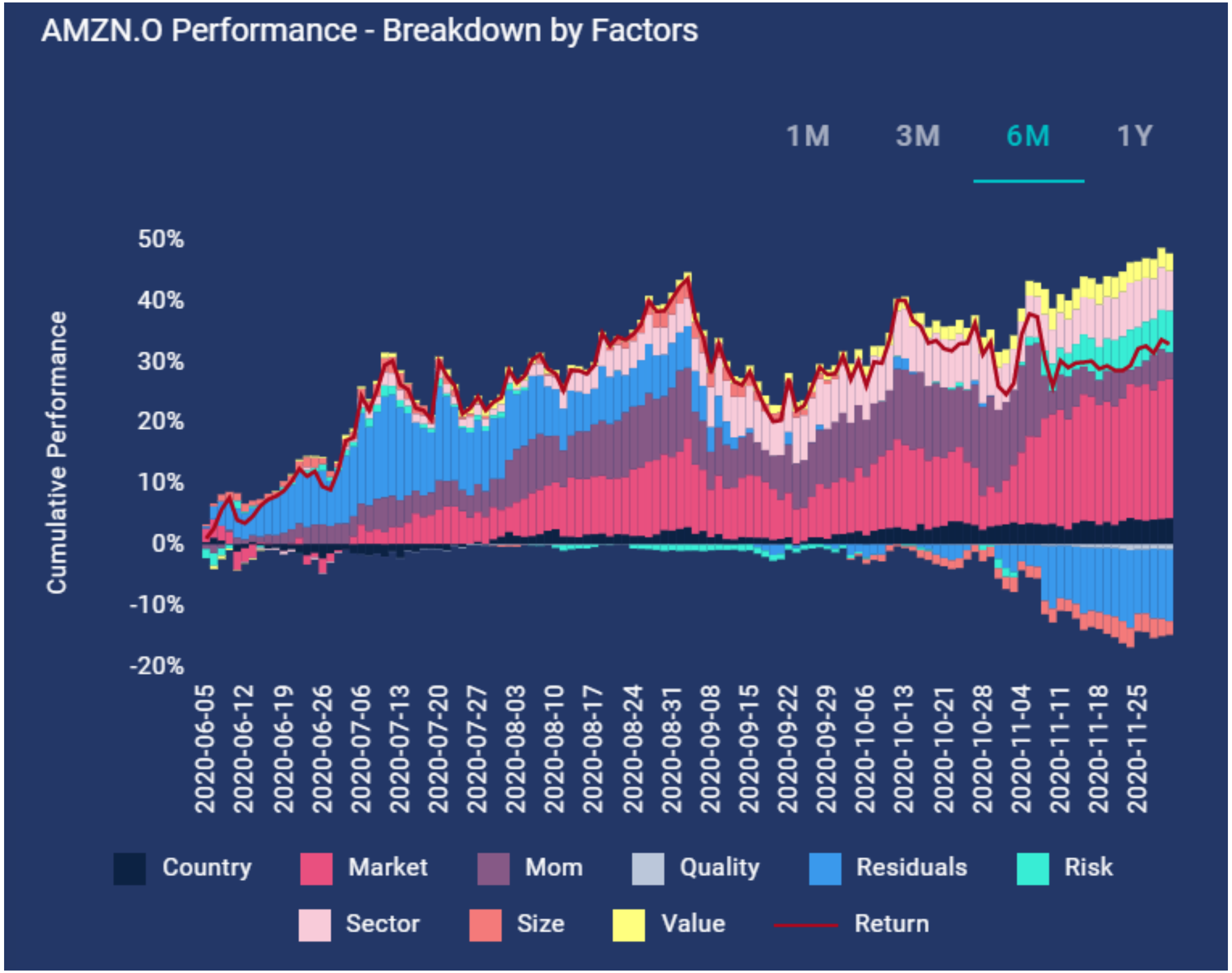
Exposure of factor k to stock n at time t

Factor returns to infer

Residuals to minimize

$$
\begin{aligned}
R_{s_1} &= X_{s_1 f_1} F_{f_1} + X_{s_1 f_2} F_{f_2} + \cdots + X_{s_1 f_{100}} F_{f_{100}} + \varepsilon_{s_1} \\
R_{s_2} &= X_{s_2 f_1} F_{f_1} + X_{s_2 f_2} F_{f_2} + \cdots + X_{s_2 f_{100}} F_{f_{100}} + \varepsilon_{s_2} \\
& \cdots \\
R_{s_{2999}} &= X_{s_{2999} f_1} F_{f_1} + X_{s_{2999} f_2} F_{f_2} + \cdots + X_{s_{2999} f_{100}} F_{f_{100}} + \varepsilon_{s_{2999}} \\
R_{s_{3000}} &= X_{s_{3000} f_1} F_{f_1} + X_{s_{3000} f_2} F_{f_2} + \cdots + X_{s_{3000} f_{100}} F_{f_{100}} + \varepsilon_{s_{3000}}
\end{aligned}
$$

Factor Model: Cross-sectional regression



Connections between Asset Pricing and Factor Models

Firm characteristics (e.g. market capitalization, accounting ratios)

$\mathrm{X}_{t,n}$

Connections between Asset Pricing and Factor Models

Firm characteristics (e.g. market capitalization, accounting ratios)

Future return $\boldsymbol{r}_{t+1,n} = f(\mathrm{x}_{t,n}) + \varepsilon_{t+1,n}$

Model
(e,g, linear model, neural network)

Connections between Asset Pricing and Factor Models

- **Factor Model: Cross-sectional regression (Explanation)**

Return of Factor k at time t (Common variables across different n)

Return $\boldsymbol{r_{t,n}} = \Sigma \mathrm{x}_{t,n,k} f_{t,k} + \varepsilon_{t,n}$

Exposure of factor k to stock n

- **Asset Pricing Model (Prediction)**

Firm characteristics: Exposure to macro-economic factors (factor loadings)

Future return $\boldsymbol{r_{t+1,n}} = f(\mathrm{x}_{t,n}) + \varepsilon_{t+1,n}$

Model
(e,g, linear model, neural network)

Empirical Asset Pricing via Machine Learning

$x_{t,n}$ Firm characteristics: Exposure to macro-economic factors (factor loadings)

Table A.6: Details of the Characteristics

No.	Acronym	Firm characteristic	Paper's author(s)	Year, Journal	Data Source	Frequency
1	absacc	Absolute accruals	Bandyopadhyay, Huang & Wirjanto	2010, WP	Compustat	Annual
2	acc	Working capital accruals	Sloan	1996, TAR	Compustat	Annual
3	aeavol	Abnormal earnings announcement volume	Lerman, Livnat & Mendenhall	2007, WP	Compustat+CRSP	Quarterly
4	age	# years since first Compustat coverage	Jiang, Lee & Zhang	2005, RAS	Compustat	Annual
5	agr	Asset growth	Cooper, Gulen & Schill	2008, JF	Compustat	Annual
6	baspread	Bid-ask spread	Amihud & Mendelson	1989, JF	CRSP	Monthly
7	beta	Beta	Fama & MacBeth	1973, JPE	CRSP	Monthly
8	betasq	Beta squared	Fama & MacBeth	1973, JPE	CRSP	Monthly
9	bm	Book-to-market	Rosenberg, Reid & Lanstein	1985, JPM	Compustat+CRSP	Annual
10	bm_ia	Industry-adjusted book to market	Asness, Porter & Stevens	2000, WP	Compustat+CRSP	Annual
11	cash	Cash holdings	Palazzo	2012, JFE	Compustat	Quarterly
12	cashdebt	Cash flow to debt	Ou & Penman	1989, JAE	Compustat	Annual
13	cashpr	Cash productivity	Chandrashekar & Rao	2009, WP	Compustat	Annual
14	cfp	Cash flow to price ratio	Desai, Rajgopal & Venkatachalam	2004, TAR	Compustat	Annual
15	cfp_ia	Industry-adjusted cash flow to price ratio	Asness, Porter & Stevens	2000, WP	Compustat	Annual
16	chatoia	Industry-adjusted change in asset turnover	Soliman	2008, TAR	Compustat	Annual
17	chcsho	Change in shares outstanding	Pontiff & Woodgate	2008, JF	Compustat	Annual
18	chempia	Industry-adjusted change in employees	Asness, Porter & Stevens	1994, WP	Compustat	Annual
19	chinv	Change in inventory	Thomas & Zhang	2002, RAS	Compustat	Annual
20	chmom	Change in 6-month momentum	Gettleman & Marks	2006, WP	CRSP	Monthly
21	chpmia	Industry-adjusted change in profit margin	Soliman	2008, TAR	Compustat	Annual
22	chtx	Change in tax expense	Thomas & Zhang	2011, JAR	Compustat	Quarterly
23	cinvest	Corporate investment	Titman, Wei & Xie	2004, JFQA	Compustat	Quarterly
24	convind	Convertible debt indicator	Valta	2016, JFQA	Compustat	Annual
25	currat	Current ratio	Ou & Penman	1989, JAE	Compustat	Annual
26	depr	Depreciation / PP&E	Holthausen & Larcker	1992, JAE	Compustat	Annual
27	divi	Dividend initiation	Michaely, Thaler & Womack	1995, JF	Compustat	Annual
28	divo	Dividend omission	Michaely, Thaler & Womack	1995, JF	Compustat	Annual
29	dolvol	Dollar trading volume	Chordia, Subrahmanyam & Anshuman	2001, JFE	CRSP	Monthly
30	dy	Dividend to price	Litzenberger & Ramaswamy	1982, JF	Compustat	Annual
31	ear	Earnings announcement return	Kishore, Brandt, Santa-Clara & Venkatachalam	2008, WP	Compustat+CRSP	Quarterly

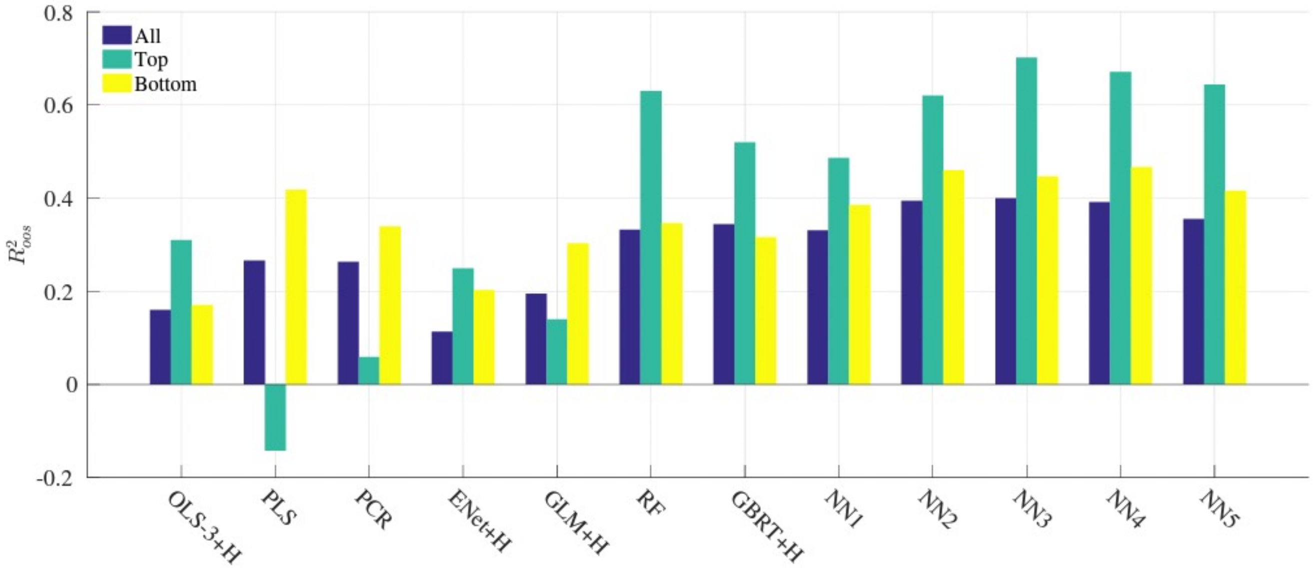
Gu, S., Kelly, B. and Xiu, D., 2020. Empirical asset pricing via machine learning. *The Review of Financial Studies*, *33*(5), pp.2223-2273.

Empirical Asset Pricing via Machine Learning

$f(\mathrm{x}_{t,n})$: $Model\ (e, g, linear\ model, neural\ network)$

Table 1: Monthly Out-of-sample Stock-level Prediction Performance (Percentage R^2_{oos})

	OLS +H	OLS-3 +H	PLS	PCR	ENet +H	GLM +H	RF	GBRT +H	NN1	NN2	NN3	NN4	NN5
All	-3.46	0.16	0.27	0.26	0.11	0.19	0.33	0.34	0.33	0.39	0.40	0.39	0.36
Top 1000	-11.28	0.31	-0.14	0.06	0.25	0.14	0.63	0.52	0.49	0.62	0.70	0.67	0.64
Bottom 1000	-1.30	0.17	0.42	0.34	0.20	0.30	0.35	0.32	0.38	0.46	0.45	0.47	0.42



Gu, S., Kelly, B. and Xiu, D., 2020. Empirical asset pricing via machine learning. *The Review of Financial Studies*, *33*(5), pp.2223-2273.

Cross-sectional Machine Learning Portfolios

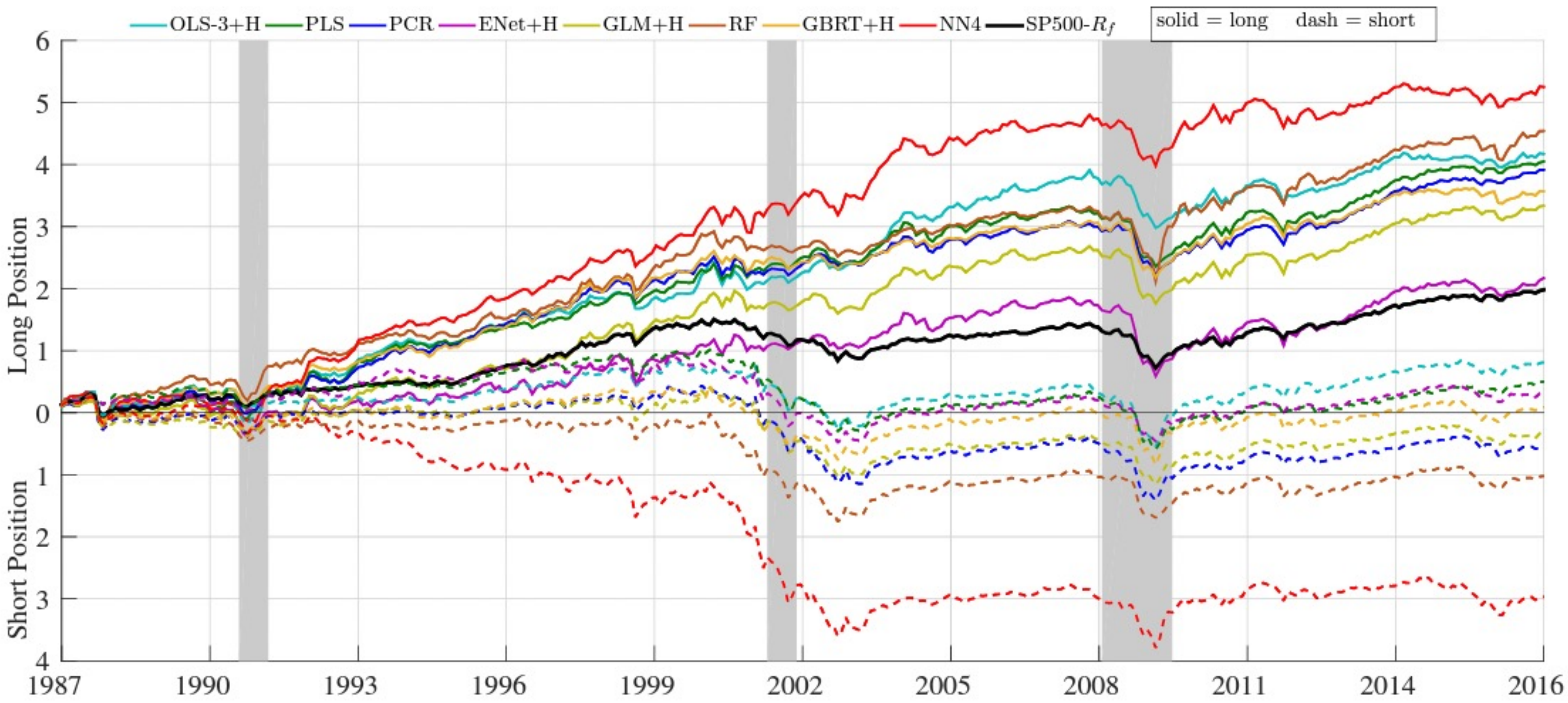
Table 7: Performance of Machine Learning Portfolios

	OLS-3+H				PLS				PCR			
	Pred	Avg	Std	SR	Pred	Avg	Std	SR	Pred	Avg	Std	SR
Low(L)	-0.17	0.40	5.90	0.24	-0.83	0.29	5.31	0.19	-0.68	0.03	5.98	0.02
2	0.17	0.58	4.65	0.43	-0.21	0.55	4.96	0.38	-0.11	0.42	5.25	0.28
3	0.35	0.60	4.43	0.47	0.12	0.64	4.63	0.48	0.19	0.53	4.94	0.37
4	0.49	0.71	4.32	0.57	0.38	0.78	4.30	0.63	0.42	0.68	4.64	0.51
5	0.62	0.79	4.57	0.60	0.61	0.77	4.53	0.59	0.62	0.81	4.66	0.60
6	0.75	0.92	5.03	0.63	0.84	0.88	4.78	0.64	0.81	0.81	4.58	0.61
7	0.88	0.85	5.18	0.57	1.06	0.92	4.89	0.65	1.01	0.87	4.72	0.64
8	1.02	0.86	5.29	0.56	1.32	0.92	5.14	0.62	1.23	1.01	4.77	0.73
9	1.21	1.18	5.47	0.75	1.66	1.15	5.24	0.76	1.52	1.20	4.88	0.86
High(H)	1.51	1.34	5.88	0.79	2.25	1.30	5.85	0.77	2.02	1.25	5.60	0.77
H-L	1.67	0.94	5.33	0.61	3.09	1.02	4.88	0.72	2.70	1.22	4.82	0.88

Gu, S., Kelly, B. and Xiu, D., 2020. Empirical asset pricing via machine learning. *The Review of Financial Studies*, *33*(5), pp.2223-2273.

Machine Learning Portfolios

Figure 9: Cumulative Return of Machine Learning Portfolios



Note: Cumulative log returns of portfolios sorted on out-of-sample machine learning return forecasts. The solid and dash lines represent long (top decile) and short (bottom decile) positions, respectively. The shaded periods show NBER recession dates. All portfolios are value weighted.

Gu, S., Kelly, B. and Xiu, D., 2020. Empirical asset pricing via machine learning. *The Review of Financial Studies*, *33*(5), pp.2223-2273.

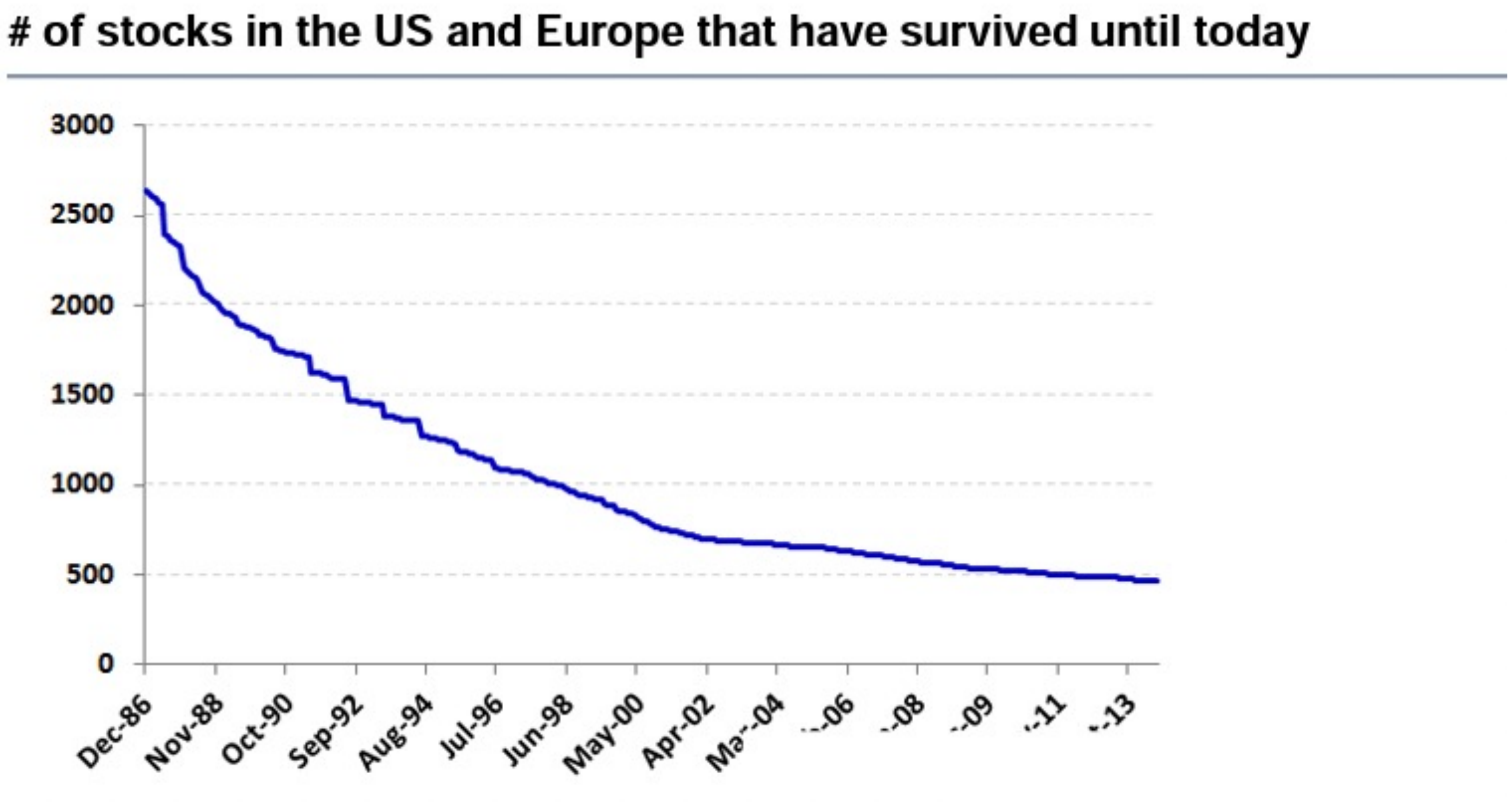
Which Characteristics Matter

Figure 5: Characteristic Importance

- Not all factor exposures are expected to earn a return premium over the long term
- Factor momentum, factor timing
- How do we know whether the factor is believable?

Gu, S., Kelly, B. and Xiu, D., 2020. Empirical asset pricing via machine learning. *The Review of Financial Studies*, *33*(5), pp.2223-2273.

Robustness Check1. Survivorship bias



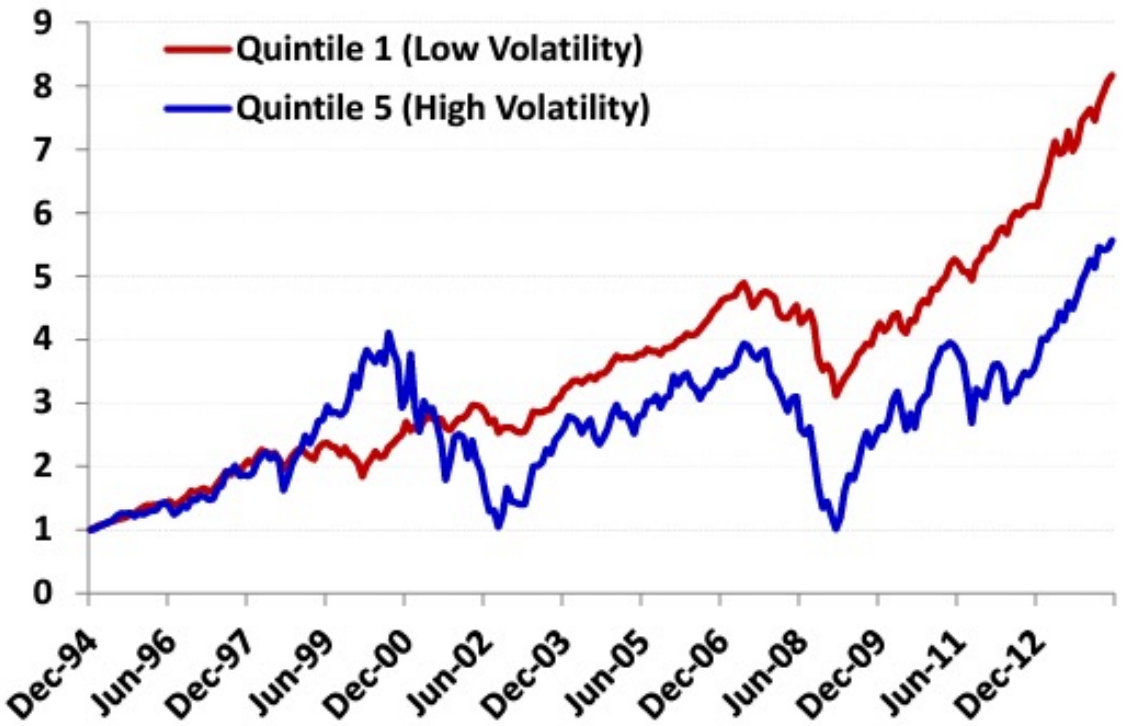
https://newyork.qwafafew.org/wp-content/uploads/sites/4/2015/10/Luo_20150128.pdf

Robustness Check1. Survivorship bias

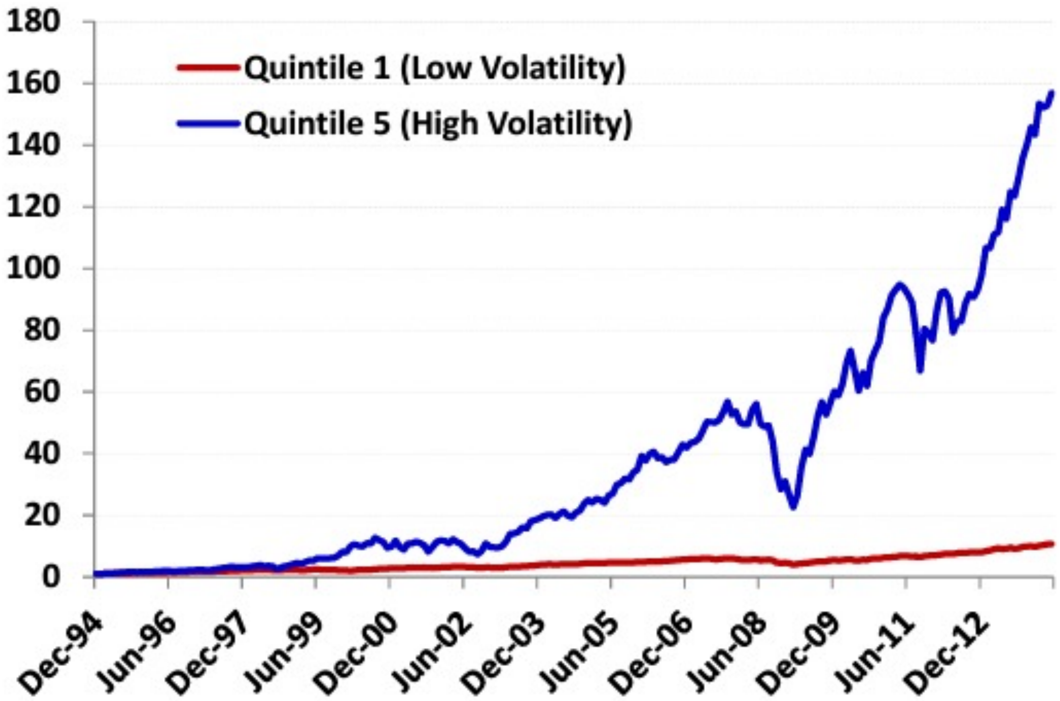
https://newyork.qwafafew.org/wp-content/uploads/sites/4/2015/10/Luo_20150128.pdf

Robustness Check1. Survivorship bias

Low volatility factor on the proper S&P 500 universe

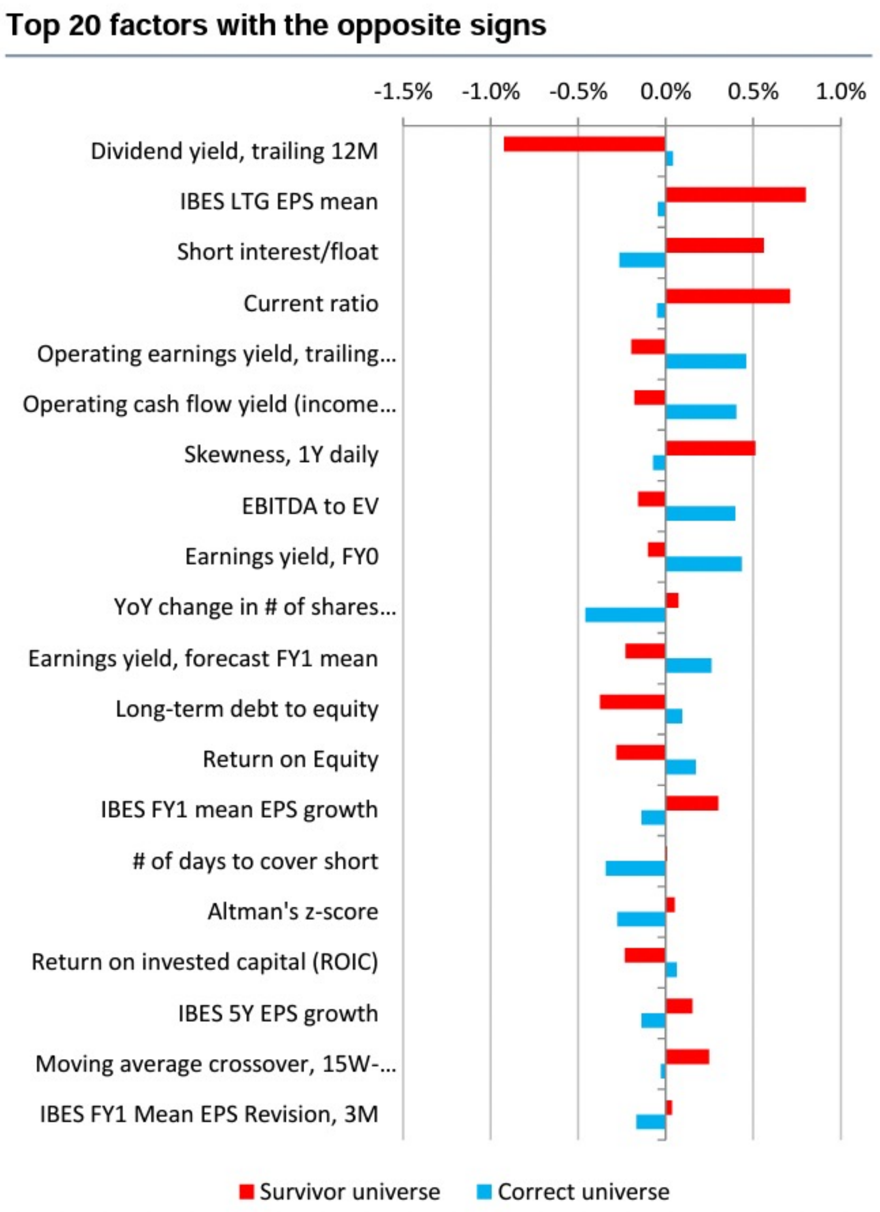


Low volatility factor performance on the current S&P 500 index constituents



https://newyork.qwafafew.org/wp-content/uploads/sites/4/2015/10/Luo_20150128.pdf

Robustness Check1. Survivorship bias

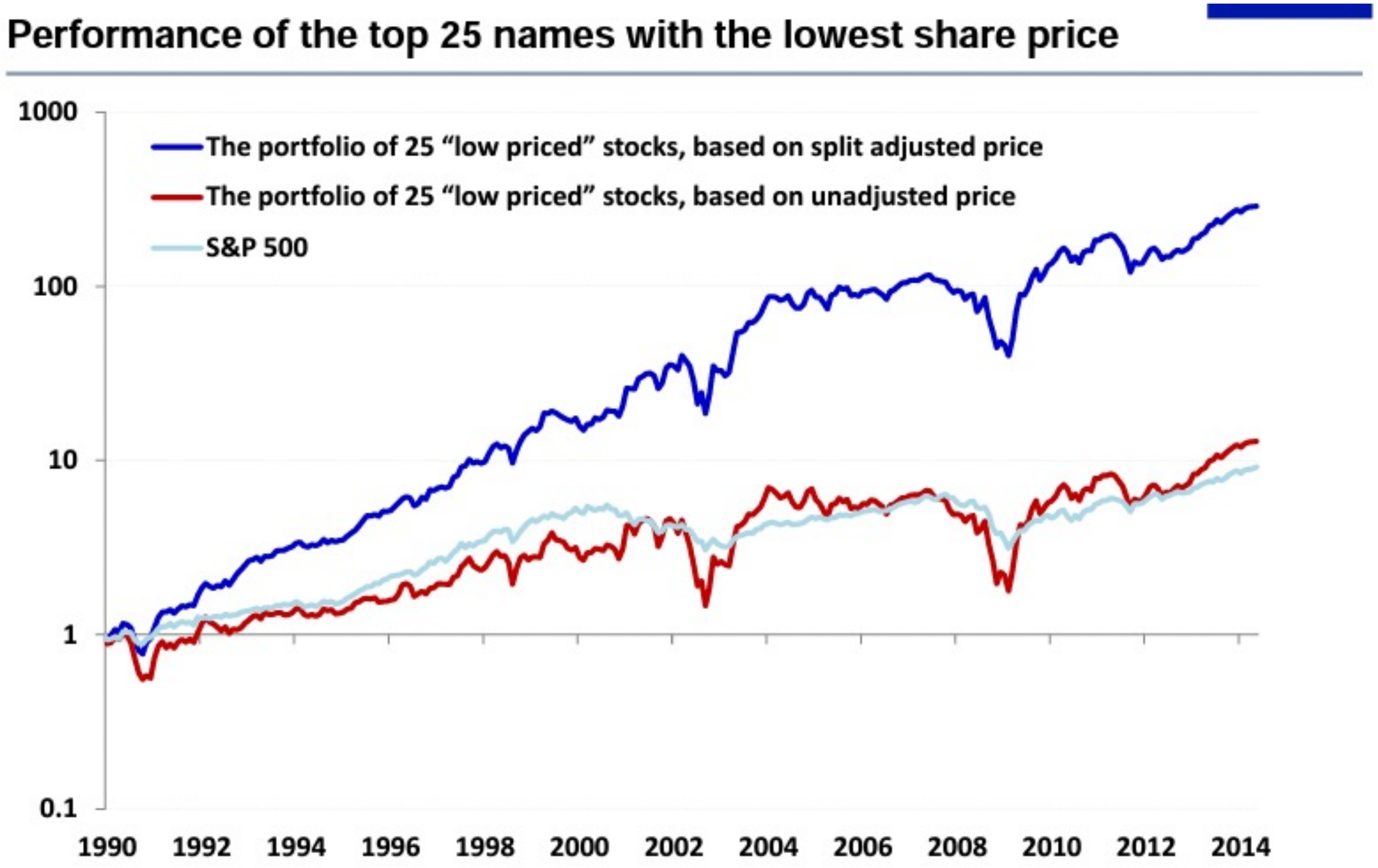


- 1/3 of factors have the opposite signs

https://newyork.qwafafew.org/wp-content/uploads/sites/4/2015/10/Luo_20150128.pdf

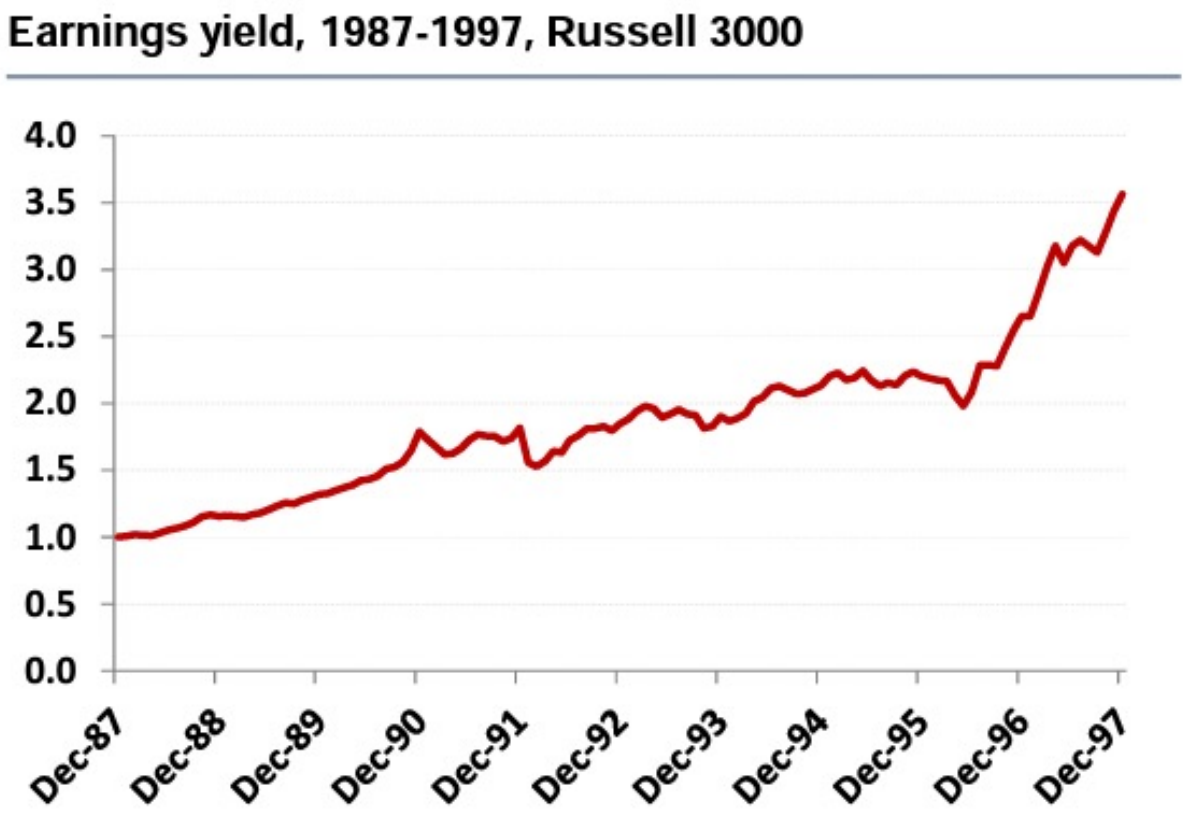
Robustness Check2. Look-ahead bias

- Using data that were unknown



https://newyork.qwafafew.org/wp-content/uploads/sites/4/2015/10/Luo_20150128.pdf

Robustness Check3. How long is long enough?

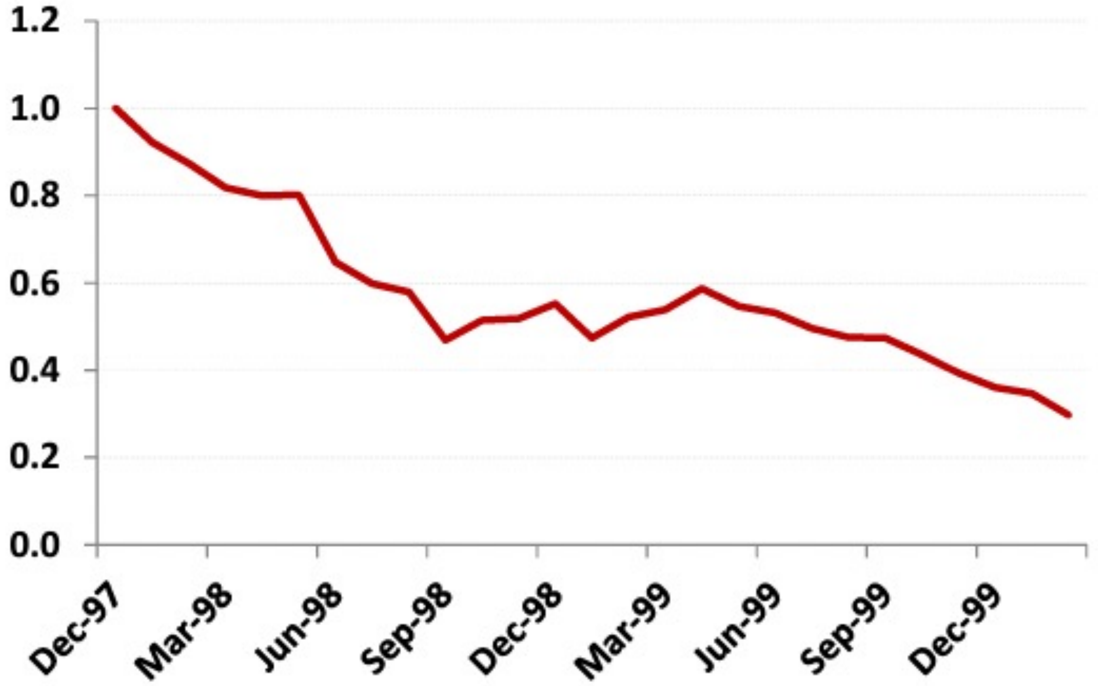


https://newyork.qwafafew.org/wp-content/uploads/sites/4/2015/10/Luo_20150128.pdf

Robustness Check3. How long is long enough?

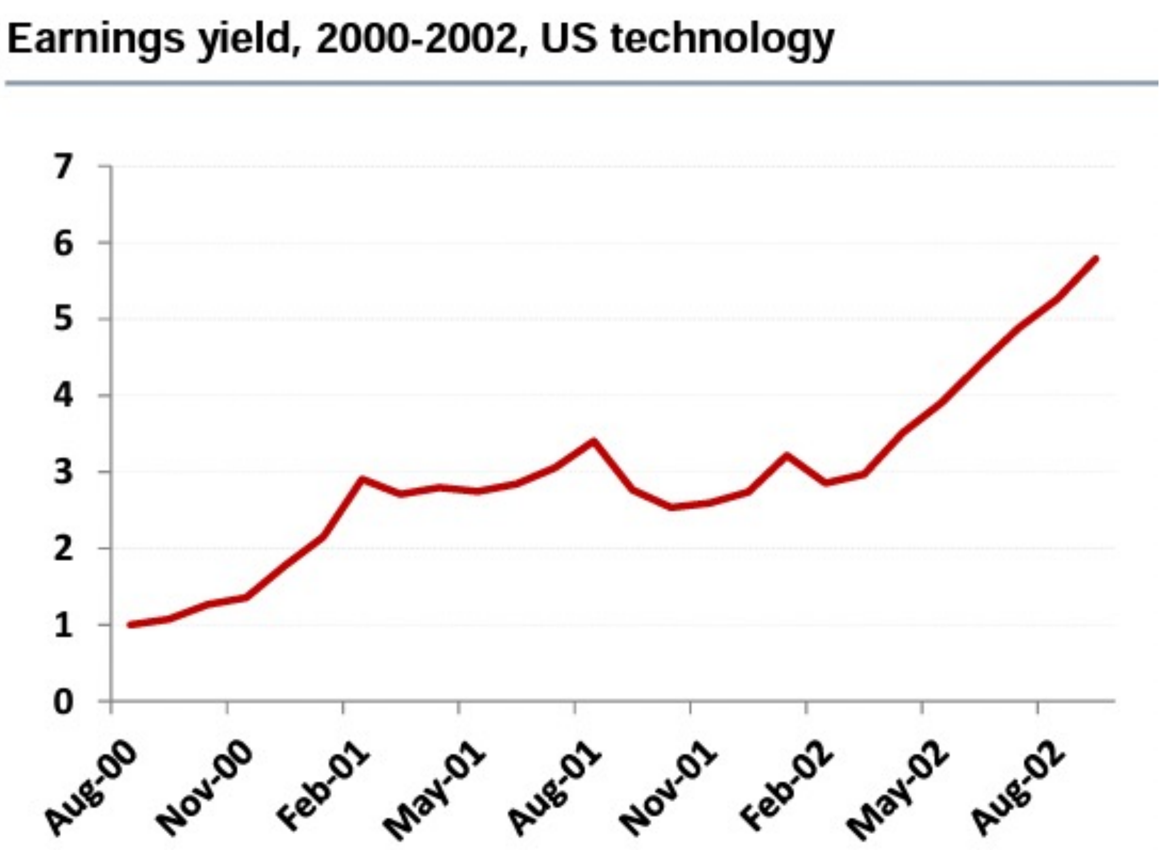
Earnings yield, 1987-1997, Russell 3000

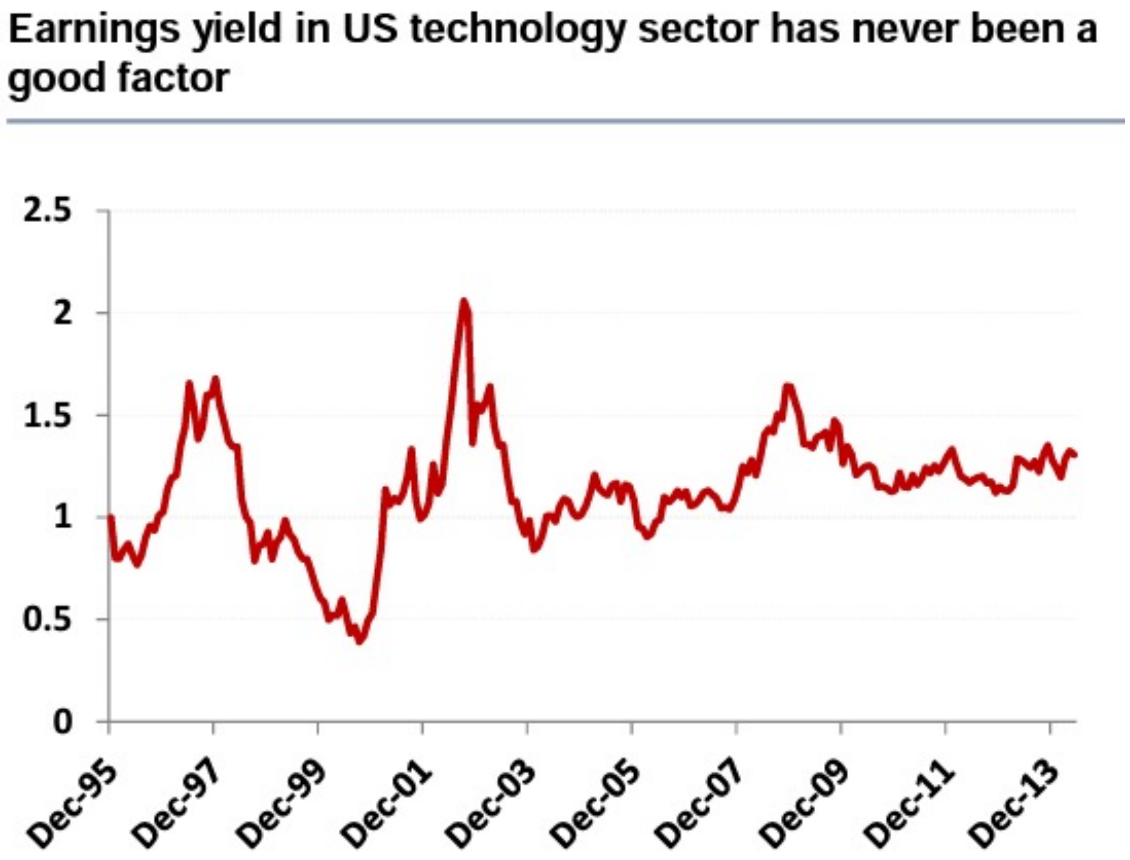
Earnings yield, 1997-2000, US technology



https://newyork.qwafafew.org/wp-content/uploads/sites/4/2015/10/Luo_20150128.pdf

Robustness Check3. How long is long enough?

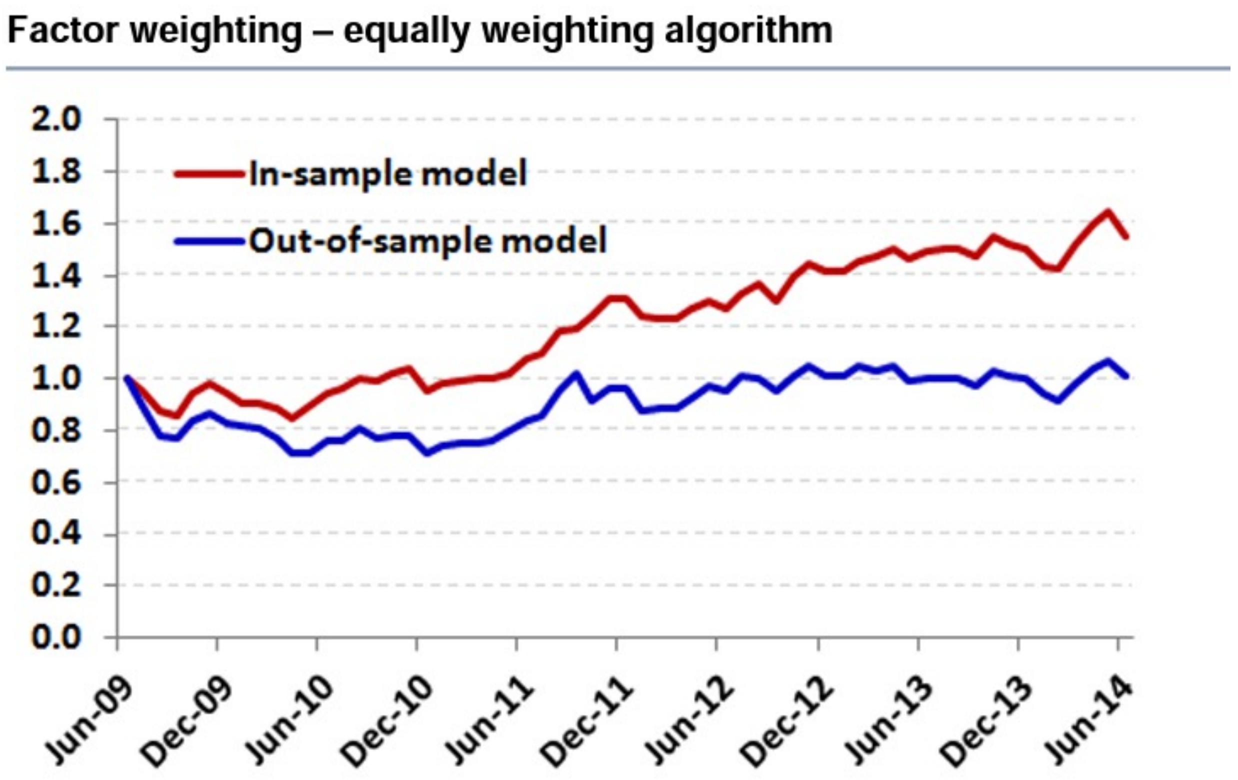




https://newyork.qwafafew.org/wp-content/uploads/sites/4/2015/10/Luo_20150128.pdf

Robustness Check4. Data mining is almost avoidable

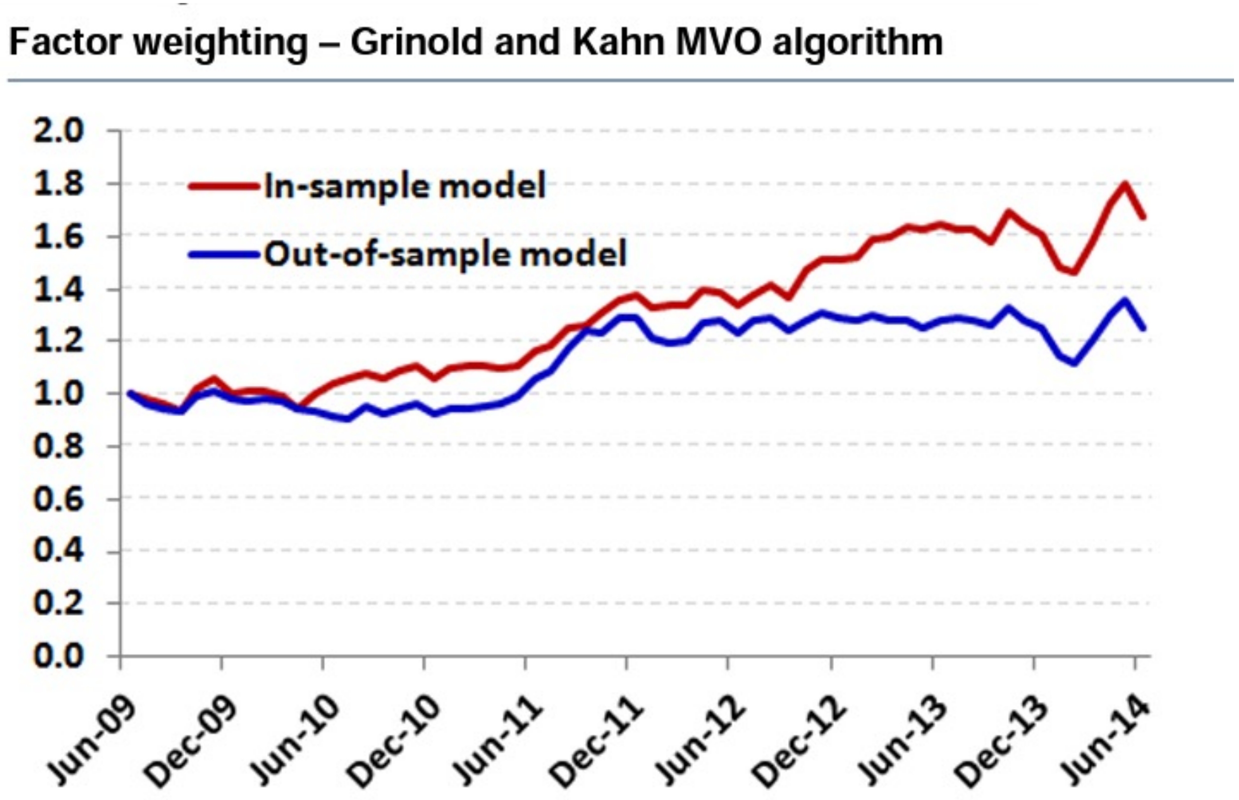
- Two factors weighting algorithms
  - Fixed backtest period; equal weights



https://newyork.qwafafew.org/wp-content/uploads/sites/4/2015/10/Luo_20150128.pdf

Robustness Check4. Data mining is almost avoidable

- Two factors weighting algorithms
  - Rolling 60 months



https://newyork.qwafafew.org/wp-content/uploads/sites/4/2015/10/Luo_20150128.pdf

Robustness Check5. The publication bias towards positive results

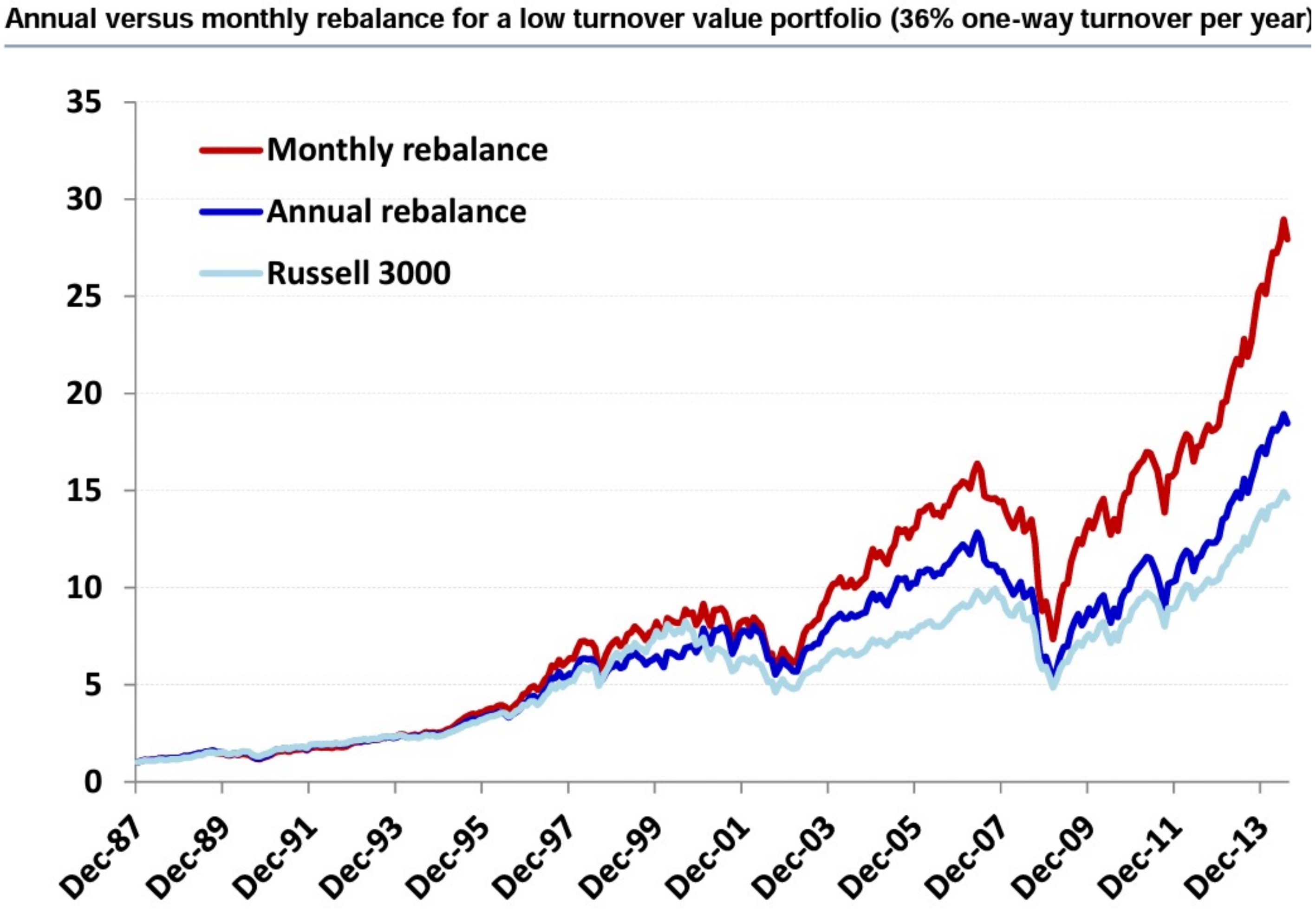
- The need for replication is therefore high and many findings have no tomorrow, especially if transaction costs are taken into account (Patton and Weller (2020), A. Y. Chen and Velikov (2020)).

Robustness Check6. The anomaly becomes public after publication

- Then, agents invest in it, which pushes prices up and the anomaly disappears

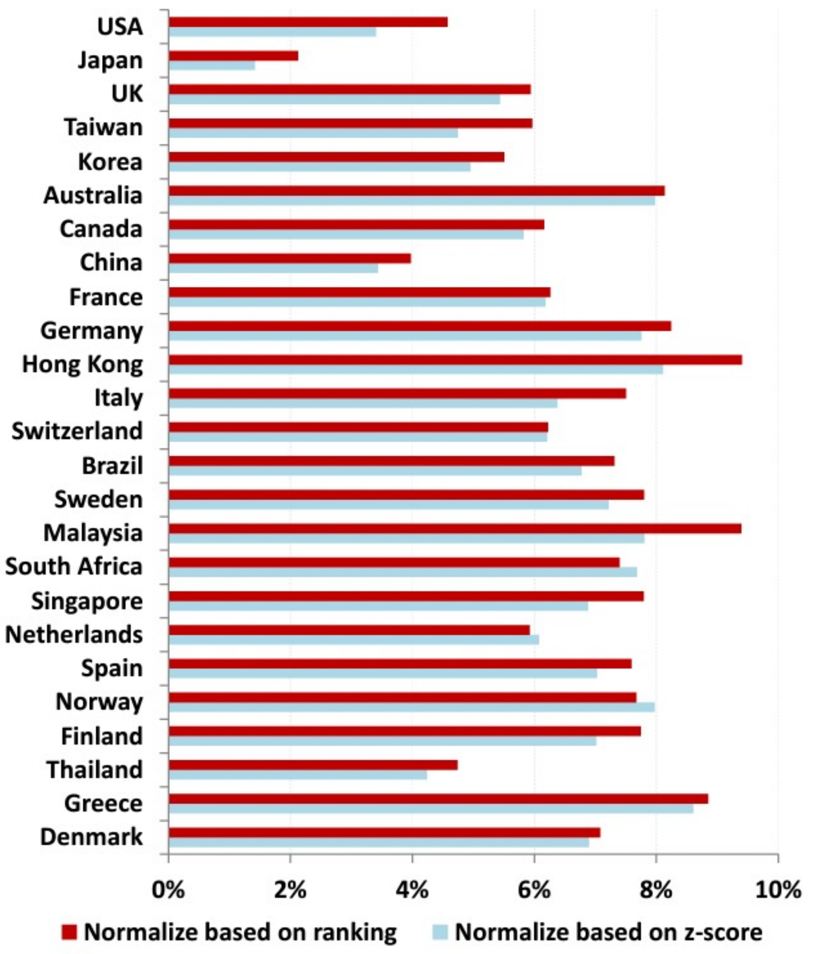
- McLean and Pontiff (2016) and Shanaev and Ghimire (2020) document this effect in the US but H. Jacobs and Müller (2020) find that all other countries experience sustained post-publication factor returns

Robustness Check7. Optimal rebalancing period

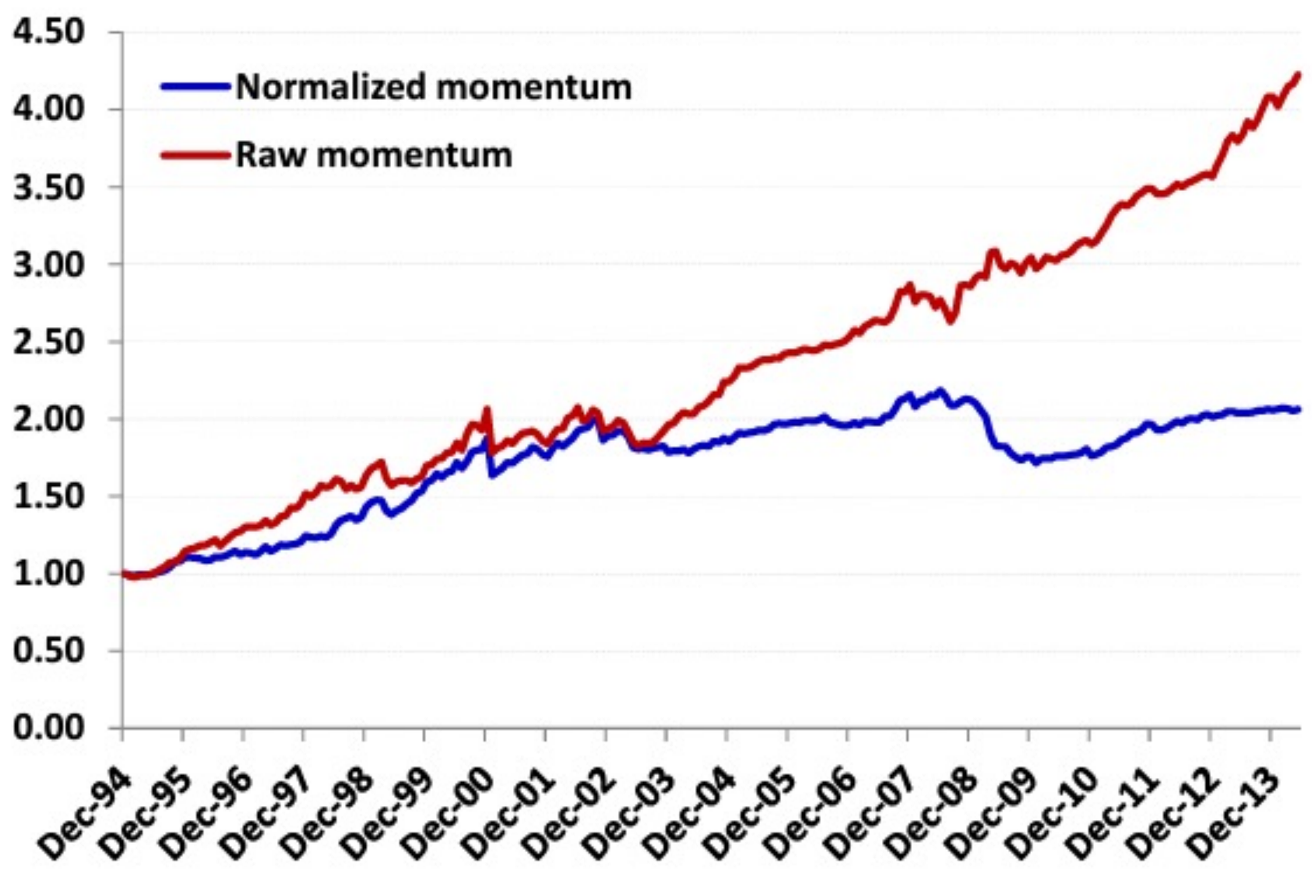


Signal decays, outliers, data transformation

Average model performance (rank IC), using different data normalization techniques



Momentum portfolio performance



Smart-beta products (ETFs)

- The democratization of so-called smart-beta products that allow investors to directly invest in particular styles (value, low volatility, etc.)

Summary of Factor Models

- The workflow of factor models
- The characteristics of good factors
- The connection between factor investing and asset pricing
- How to implement factor exposures to your portfolio
  - Long only, Long/Short position based on a quantile portfolio
  - Factor-tilts
  - Smart beta (ETFs)

Additional Materials

- Your investment skills are proportional to the reading volume of finance literature & backtesting practices

- But, not all papers are meaningful
  - **Academic Journals:** the Journal of Finance, the Review of Financial Studies, the Journal of Financial Economics
  - **Practitioner Journals:** the Journal of Portfolio Management, the Financial Analysts Journal
  - **Working Papers (SSRN):** Why? Market efficiency, Decaying alphas

Thank You

References

- Kenneth French, https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html
- Seven Sins of Backtesting, https://newyork.qwafafew.org/wp-content/uploads/sites/4/2015/10/Luo_20150128.pdf
- ML Factors, http://www.mlfactor.com/
- Equity factor-based investing: A practitioner’s guide, https://www.vanguardinvestments.de/documents/institutional/factors-whitepaper-eu.pdf
- Do Portfolio Factors or Characteristics Drive Expected Returns?, https://alphaarchitect.com/2017/10/factors-vs-characteristics/
- Equity Factor Models - Build one in R with a few lines of codes, https://towardsdatascience.com/custom-factor-models-build-your-own-in-r-with-a-few-lines-of-codes-502274ae3624