FE670 Algorithmic Trading Strategies

Lecture 5. Cross-Sectional Models, Factor Zoo, and Factor Trading Strategies

Steve Yang

Stevens Institute of Technology

steve.yang@stevens.edu

10/05/2023

Overview

Desirable Properties of Factors

Building Factors from Company Characteristics

Multi-Factor Portfolios

Barra Factor Analysis

Performance Evaluation of Factors - Factor Zoo

Methodologies for a Factor-Based Trading Strategy

Python Sample - Fundamental Factor Long Short Strategy

Desirable Properties of Factors

Factors should be founded on sound economic intuition, market insight, or an anomaly. In addition to the underlying economic reasoning, factors should have other properties that make them effective for forecasting:

- It is an advantage if factors are **intuitive** to investors. Many investors will only invest in particular funds if they understand and agree with the basic ideas behind the trading strategies. Factors give portfolio managers a tool in communicating to investors what themes they are investing in.
- The search for the **economic meaningful** factors should avoid strictly relying on pure historical analysis. Factors used in a model should not emerge from a sequential process of evaluating successful factors while removing less favorable ones.

- A group of factors should be **parsimonious** in its description of the trading strategy. This will require careful evaluation of the interaction between the different factors. For example, highly correlated factors will cause the interferences made in a multivariate approach to be less reliable.

- The success or failure of factors selected should not depend on a few outliers. It is desirable to construct factors that are reasonably **robust to outliers**.

Source of Factors: The sources are widespread with no one dominating clearly. Search through a variety of sources seems to provide the best opportunity to uncover factors that will be valuable for new models. Example sources include economic foundations, inefficiency in processing information, financial reports, discussions with portfolio managers or traders, sell-side reports or equity research reports, and academic literature in finance, accounting, and economics, etc.

Building Factors from Company Characteristics

- We desire our factors to relate the financial data provided by a company to metrics that investors use when making decisions about the attractiveness of a stock such as valuation ratios, operating efficiency ratios, profitability ratios, and solvency ratios.

- Factors should also relate to the market data such as forecasts, prices and returns, and trading volume. We distinguish three categories of financial data: time series, cross-sectional, and panel data.

- *Time series* data consist of information and variables collected over multiple time periods. *Cross-sectional* data consist of data collected at one point in time for many different companies. *A panel* data set consists of cross-sectional data collected at different points in time.

Data Integrity

Quality data maintain several attributes such as providing a consistent view of history, maintaining good data availability, containing no survivorship, and avoid look-ahead bias. It is important for the quantitative researchers to be able to recognize the limitations and adjust the data accordingly.

1. *Backfilling* of data happens when a company is first entered into a database at the current period and its historical data are also added.
2. *Restatement* of data are prevalent in distorting consistency of data. Many database companies may overwrite the number initially recorded.
3. *Survivorship bias* occurs when companies are removed from the database when they no longer exist.
4. *Lookahead bias* occurs when data are used in a study that would not have been available during the actual period analyzed.

Multi-Factor Portfolios

- Factor portfolios are constructed to measure the information content of a factor. The objective is to mimic the return behavior of a factor and minimize the residual risk. Similar to portfolio sorts, we evaluate the behavior of these factor portfolios to determine whether a factor earns a systematic premium.

- Typically, a factor portfolio has a unit exposure to a factor and zero exposure to other factors. Construction of factor portfolios requires holding both long and short positions. We can also build a factor portfolio that has exposure to multiple attributes, such as beta, sectors, or other characteristics. Portfolios with exposures to multiple factors provide the opportunity to analyze the interaction of different factors.

A Factor Model Approach

- By using a multifactor model, we can build factor portfolios that control for different risks. We decompose return and risk at a point in time into a systematic and specific component using the regression:

  $$\mathbf{r} = \mathbf{X}\mathbf{b} + \mathbf{u}$$

  where **r** is an N vector of excess returns of the stocks considered, **X** is an N by K matrix of factor loadings, **b** is a K vector of factor returns, and **u** is a N vector of firm specific returns (residual returns).
- We assume that factor returns are uncorrelated with the firm specific return. Further assuming that firm specific returns of different companies are uncorrelated, the N by N covariance matrix of stock return **V** is given by:

  $$\mathbf{V} = \mathbf{X}\mathbf{F}\mathbf{X}' + \boldsymbol{\Delta}$$

where $\mathbf{F}$ is the K by K factor return covariance matrix and $\mathbf{\Delta}$ is the N by N diagonal matrix of variances of the specific returns.

- We can use the Fama-MacBeth procedure discussed earlier to estimate the factor returns over time. Each month, we perform Generalize Least Square - GLS regression to obtain

$$\mathbf{b} = (\mathbf{X}'\mathbf{\Delta}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Delta}^{-1}\mathbf{r}$$

OLS would give us an unbiased estimate, but since the residuals are heteroskedastic the GLS methodology is preferred and will deliver a more efficient estimate. The resulting holdings for each factor portfolio are given by the rows of $(X'\Delta^{-1}X)^{-1}X'\Delta^{-1}$.

An Optimization-Based Approach

- A second approach to build factor portfolios uses mean-variance optimization. Using optimization techniques provide a flexible approach for implementing additional objectives and constrains. We would like to construct a portfolio that has maximum exposure to one targeted factor from $\mathbf{X}$ (the alpha factor), zero exposure to all other factors, and minimum portfolio risk. Let us denote the alpha factor by $\mathbf{X}_\alpha$ and all the remaining ones by $\mathbf{X}_\sigma$. Then the resulting optimization problem takes the form:

$$\max_{w} \left\{ \mathbf{w}'\mathbf{X}_\alpha - \frac{\mathbf{1}}{\mathbf{2}}\lambda \mathbf{w}'\mathbf{V}\mathbf{w} \right\}$$

$$s.t. \mathbf{w}'\mathbf{X}_\sigma = 0$$

- The analytical solution is given by:

$$h^* = \frac{\mathbf{1}}{\lambda}\mathbf{V}^{-\mathbf{1}}\left[\mathbf{I} - \mathbf{X}_\sigma(\mathbf{X}_\sigma'\mathbf{V}^{-\mathbf{1}}\mathbf{X}_\sigma)^{-\mathbf{1}}\mathbf{X}_\sigma'\mathbf{V}^{-\mathbf{1}}\right]\mathbf{X}_\alpha$$

Methods to Adjust Factors

- A factor may need to be adjusted using analytical or statistical techniques to be more useful for modeling. The following three adjustment are common:
- **Standardization:** It rescales a variable while preserving its order. Typically, we choose the standardized variable to have a mean of zero and a standard deviation of one by using the transformation

$$x_i^{new} = \frac{x_i - \bar{x}_i}{\sigma_x}$$

- **Orthogonalization:** Orthogonalizing a factor for other specified factor(s) removes this relationship. To orthogonalize the factor using averages according to industries or sectors, we can first calculate industry scores

$$s_k = \frac{\sum_{i=1}^{n} x_i \cdot \mathrm{ind}_{i,k}}{\sum_{i=1}^{n} \mathrm{ind}_{i,k}}$$

where x_i is a factor and $\text{ind}_{i,k}$ represent the weight of stock i in industry k. Next we subtract the industry average of the industry scores, s_k, from each stock. We compute

$$x_i^{new} = x_i - \sum_{k \in \text{Industries}} \text{ind}_{i,k} \cdot s_k$$

where x_i^{new} is the new industry neutral factor.

We can also use linear regression to orthogonalize a factor. We first determine the coefficients in the equation

$$x_i = a + b \cdot f_i + \epsilon_i$$

where f_i is the factor to orthogonalize the factor x_i by, b is the contribution of f_i to x_i, and ϵ_i is the component of the factor x_i not related to $f_i \cdot \epsilon_i$ is orthogonal to f_i (that is, ϵ_i is independent of f_i) and represents the neutralized factor $x^{\text{new}} = \epsilon_i$

In the same fashion, we can orthogonalize our variable relative to a set of factors by using the multivariate linear regression

$$x_i = a + \sum_j b_j \cdot f_j + \epsilon_i$$

and then setting $x_j^{new} = \epsilon_i$.

The interaction between factors in a risk model and an alpha model often concerns portfolio managers. One possible approach to address this concern is to orthogonalize the factors or final scores from the alpha model against the factors used in the risk model.

- **Transformation:** It is a common practice to apply transformations to data used in statistical and econometric models. In particular, factors are often transformed such that the resulting series is symmetric or close to being normally distributed. Frequently used transformations include natural logarithms, exponentials, and square roots.

- **Outliers Detection and Management:** Outliers are observations that seem to be inconsistent with the other values in a data set. Financial data contain outliers for a number of reasons including data errors, measurement errors, or unusual events.

- Outliers can be detected by several methods. Graphs such as boxplots, scatter plots, or histograms can be useful to visually identify them. Alternatively there are number of numerical techniques available. One common method is to compute the inter-quantile-range and then identify outliers as a measure of dispersion and is calculated as the difference between the third and first quartiles of a sample.

- Winsorization is the process of transforming extreme values in the data. First, we calculate percentiles of the data. Next we define outliers by referencing a certain percentile ranking. It is important to fully investigate the practical consequences of using either one of these procedures.

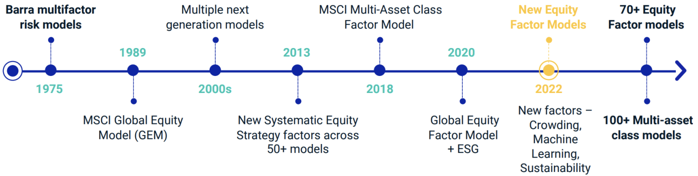
Barra Factor Analysis

- Barra factor models are based on the multi-factor models, a concept Barra first developed in 1975. The model setup is as follows: given N asset prices, and K fundamental factor loadings at time t, the Barra fundamental factor model express a linear relationship:

$$E[r_t] = \mathbf{F}_{\theta_t}(\mathbf{B}_t) + \epsilon_t = \mathbf{B}_t\theta_t + \epsilon_t, t = 1, ..., T. \tag{1}$$

  where $\mathbf{B}_t$ is the $NxK + 1$ matrix of known factor loadings. The term $\theta_t = [\alpha_t, f_{1,t}, ..., f_{K,t}]$ is the $K + 1$ vector of unobserved factor realizations at time t.

MSCI Model Evolution



1975 **Barra multifactor risk models**

1989 MSCI Global Equity Model (GEM)

2000s Multiple next generation models

2013 New Systematic Equity Strategy factors across 50+ models

2018 MSCI Multi-Asset Class Factor Model

2020 Global Equity Factor Model + ESG

2022 **New Equity Factor Models**

New factors – Crowding, Machine Learning, Sustainability

70+ Equity Factor models

100+ Multi-asset class models

Barra Factor Analysis

Compare with Factor Pricing Model

Consider an economy with N^t stocks at each time $t = 1, \cdot, T$. Let $r_{i,t}$ denote the return for the i-th firm at time t. Assume there are J risk factors.

$$r_{i,t+1} = \alpha_{i,t} + \beta'_{i,t} f_{t+1} + \epsilon_{i,t+1} \tag{2}$$

where $\mathbb{E}_t[\epsilon_{i,t+1}] = \mathbb{E}_t[\epsilon_{i,t+1} f_{t+1}] = 0$, $\mathbb{E}_t[f_{t+1}] = \lambda_t$, and $\alpha_{i,t} = 0$ for all i and t, $\beta_{i,t}, f_{t+1} \in \mathbb{R}^J$ are J-dimensional vector.

Conditional Expected Return

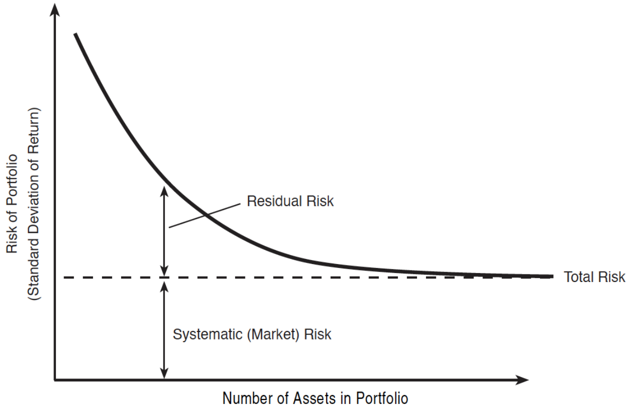
$$\mathbb{E}_t[r_{i,t+1}] = \beta'_{i,t}\lambda_t = \sum_{j=1}^{J} \beta_{i,t}^{(j)} \lambda_t^{(j)} \tag{3}$$

Barra Factor Analysis

- A central part of the model is its factor covariance matrix. This matrix contains the variances and covariances of the common factors. To estimate a portfolio's risk, we must consider not only the security or portfolio's exposures to the factors, but also each factor's risk and the covariance or interaction between factors.
- Without the framework of a multiple-factor model, estimating the covariance of each asset with every other asset would likely result in finding spurious relationships. For example, an estimation universe of 1,400 assets entails 980,700 covariances and variances to calculate.
- Barra's risk models use historical returns to create a framework for predicting the future return volatility of an asset or a portfolio. Each month, the estimation universe, which is the set of representative assets in each local market, is used to attribute asset returns to common factors and to a specific, or residual, return.

Barra Factor Analysis

- Financial theorists became more scientific and statistical in the early 1950s. Harry Markowitz was the first to quantify risk (as standard deviation) and diversification. In the late 1950s, Leo Breiman and John L. Kelly Jr. derived mathematically the peril of ignoring risk.



Prediction with Barra Factor

- Prediction involves applying the Barra model, trained in the previous period, to a new observation $\mathbf{B}_t$. The fitted map $\mathbf{F}_{\hat{\theta}_{t-1}}(\mathbf{B}_t) = \mathbf{B}_t \theta\hat{}_{t-1}$ give the next period returns

$$\hat{\mathbf{r}}_{t+1} = \mathbf{F}_{\hat{\theta}_{t-1}}(\mathbf{B}_t), t = 1, ..., T-1. \tag{4}$$

- Each supervised training set is a factor loading matrix $\mathbf{B}_t$ and asset return N-vector $\mathbf{r}_t$. Similarly, each supervised test set is the pair $(\mathbf{B}_{t+1}, \mathbf{r}_t)$.
- The form of the statistical model (4) uses point-wise estimation of each asset return from its $\mathbf{K}$ factor loadings. Our goal is to cast the Barra model into a Bayesian deep learning framework.

Bayesian Deep Learning Model

- Given i.i.d. data $\mathcal{D}_t := (\mathbf{X}, \mathbf{Y}) = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n$, deep feedforward neural networks learn a parameter forward map (Polson and Sokolov, 2017)
$$\hat{\mathbf{Y}} = \mathbf{F}_\theta(\mathbf{X}), \tag{5}$$
where $\mathbf{F}_\theta$ is a superposition of L univariate *semi-affine* functions, $\sigma_{\theta^\ell}^{(\ell)}, \ell \in \{1, ..., L\}$:
$$\mathbf{F}_\theta(\mathbf{X}) = \left(\sigma_{\theta^L}^{(L)} \circ \cdots \circ \sigma_{\theta^1}^{(1)}\right)(\mathbf{X}), \tag{6}$$
and the unknown parameters $\theta = (\mathbf{W}, \mathbf{b})$ are a set of weight matrices $\mathbf{W} = (W^{(1)}, ..., W^{(L)})$ and $\mathbf{b} = ((b^{(1)}, ..., b^{(L)})$.
- The ℓth semi-affine function is itself defined as the composition of the activation function, $\sigma^{(\ell)}(\cdot)$ and an affine map:
$$\sigma_{\theta^\ell}^{(\ell)}(\mathbf{Z}^{(\ell-1)}) = \sigma^{(\ell)}\left(\mathbf{W}^{(\ell)}\mathbf{Z}^{(\ell-1)} + \mathbf{b}^{(\ell)}\right), \tag{7}$$

Performance Evaluation of Factors - Factor Zoo

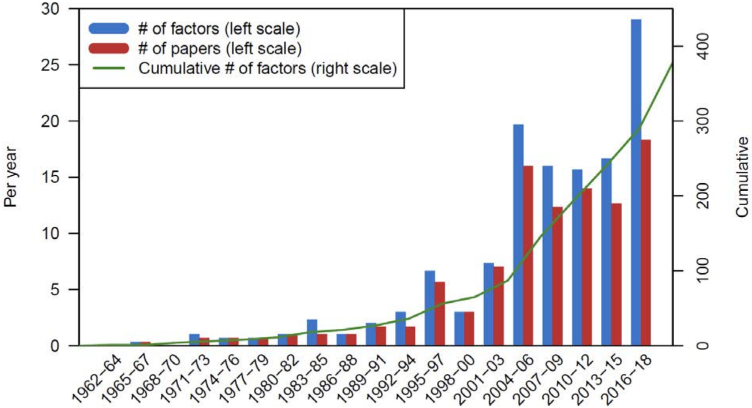
- Analyzing the performance of different factors is an important part of the development of a factor-based trading strategy.

- A researcher may construct and analyze over a hundred different factors, so that means to evaluate and compare these factors is needed. Most often this process starts by trying to understand the time-series properties of each factor in isolation and then study how they interact with each other.

- To better understand the time variation of the performance of these factors, one may calculate rolling 24-month mean returns and correlations of the factors. But more thorough statistical analyses would be required as we see the exploding number of factors proposed recently.

Factor "Zoo" - Evaluation

- An initial census of the zoo was carried out by Harvey, Liu, and Zhu (2016), who detail over 300 factors published in top academic journals based on data through 2012.



* Journals published through December 2018. Data collection in January 2019.

Issues with Factor Zoo

- First, we must take multiple testing into account when assessing statistical significance. We usually focus on an acceptable rate of false positives (e.g., a 5% level) and will often declare a factor significant at the 5% level (e.g., two standard errors from zero, aka two sigma). This works for a single try, but if we test for instance, 20 different factors, one will likely be two sigma - purely by chance. If we accept this factor as a true factor, the error rate will not be 5%, but closer to 60%.
- Second, many multiple testing corrections are suggested in the literature. The simplest is the Bonferroni correction. Suppose we try 50 factors and find that one is approximately three sigma with a p-value of 0.01. Three sigma is impressive under a single test (usually we look for a p-value < 0.05) - but we did 50 tests. The Bonferroni correction simply multiplies the p-value by the number of tests. So the Bonferroni-adjusted p-value is 0.50, which is much larger than our usual 0.05.

Issues with Factor Zoo

- **Family-wise Error Rate (FWER:** Holm (1979) is the first to formally define the family-wise error rate:
  - i). For a single hypothesis test, a value α is used to control type I error rate: the probability of finding a factor to be significant when it is not. The α is sometimes called the "level of significance." In a multiple testing framework, restricting each individual test?s type I error rate at ? is not enough to control the overall probability of false discoveries.
  - ii). The family-wise error rate (FWER) is the probability of at least one type I error: $\text{FWER} = P_r(N_{0|r}) \geq 1)$.
  - iii). FWER measures the probability of even a single false discovery, regardless of the total number of tests.
- For instance, researchers might test 100 factors; FWER measures the probability of incorrectly identifying one or more factors to be significant. Given significance or threshold level α, we explore two existing methods (Bonferroni and Holm's adjustment) to ensure FWER does not exceed α.

Issues with Factor Zoo

- **False Discovery Rate (FDR):** The false discovery proportion (FDP) is the proportion of type I errors:

$$\mathrm{FDP} = \begin{cases} \frac{N_{0|r})}{R} & \text{if} R > 0, \\ 0 & \text{if} R = 0 \end{cases}$$

  i). The false discovery rate (FDR) is defined as $\mathrm{FDR} = E[\mathrm{FDP}]$.

  ii). The family-wise error rate (FWER) is the probability of at least one type I error: $\mathrm{FWER} = P_r(N_{0|r}) \geq 1)$.

  iii). FDR measures the expected proportion of false discoveries among all discoveries. It is less stringent (i.e., leads to more discoveries) than FWER and usually much less so when many tests are performed.

  iv). Intuitively, this is because FDR allows $N_{0|r}$ to grow in proportion to R, whereas FWER measures the probability of making even a single type I error.

- The statistics literature has developed many methods to control both FWER and FDR.

Issues with Factor Zoo

- **Bonferroni's Adjustment:** Bonferroni's adjustment is as follows:
  - i). Reject any hypothesis with p-value $\frac{\alpha_\omega}{M}$:

$$p_i^{\text{Bonferroni}} = \min[M \times p_i, 1]$$

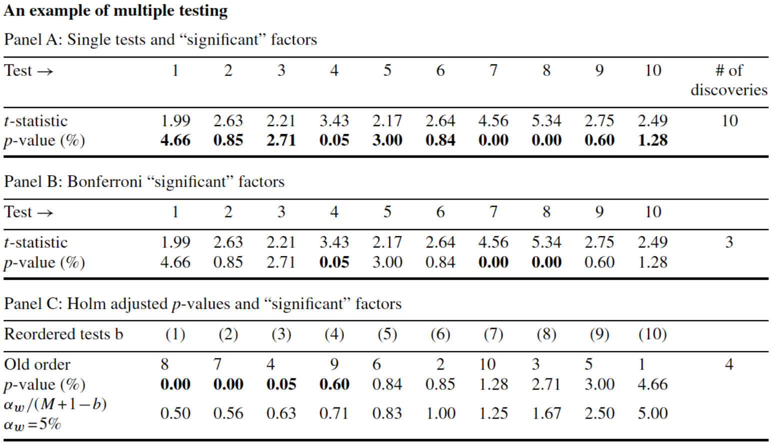
  - ii). Let k be the minimum index such that $p_{(b)} > \frac{\alpha_\omega}{M+1-b}$.
  - iii). Bonferroni applies the same adjustment to each test. It inflates the original p-value by the number of tests M
  - iv). The adjusted p-value is compared with the threshold value α_ω.
- Bonferroni operates as a single-step procedure that can be shown to restrict FWER at levels less than or equal to $(M_0 \times \alpha_\omega)/M$, without any assumption on the dependence structure of the p-values. Since $M_0 \leq M$, Bonferroni also controls FWER at level α_ω.

Issues with Factor Zoo

- **Holm's Adjustment:** Sequential methods have been proposed to adjust p-values in multiple hypothesis testing (Schweder and Spjotvoll,1982):
  i). Order the original p-values such that $p_{(1)} \leq p_{(2)}) \leq \ \leq p_{(b)}) \leq \ \leq p_{(M)}$, and let the associate null hypotheses be $H_{(1)}, H_{(2)}, ..., H_{(b)}, ..., H_{(M)}$.
  ii). Let k be the minimum index such that $p_{(b)} > \frac{\alpha_\omega}{M+1-b}$.
  iii). Reject the null hypotheses $H_{(1)}, H_{(2)}, ..., H_{(k-1)}$ (i.e., declare these factors significant), but not $H_{(k)}, ..., H_{(M)}$.
- Holm's adjustment is a step-down procedure: for the ordered p-values, we start from the smallest p-value and go down to the largest one.
- If k is the smallest index that satisfies $p_{(b)} > \frac{\alpha_\omega}{M+1-b}$, we will reject all tests whose ordered index is below k.

Issues with Factor Zoo

- Like Bonferroni, Holm also restricts FWER at α_ω without any requirement on the dependence structure of p-values (Harvey, Zhang, and Liu, 2016).



An example of multiple testing

Panel A: Single tests and "significant" factors

Test →	1	2	3	4	5	6	7	8	9	10	# of discoveries
t-statistic	1.99	2.63	2.21	3.43	2.17	2.64	4.56	5.34	2.75	2.49	10
p-value (%)	**4.66**	**0.85**	**2.71**	**0.05**	**3.00**	**0.84**	**0.00**	**0.00**	**0.60**	**1.28**	

Panel B: Bonferroni "significant" factors

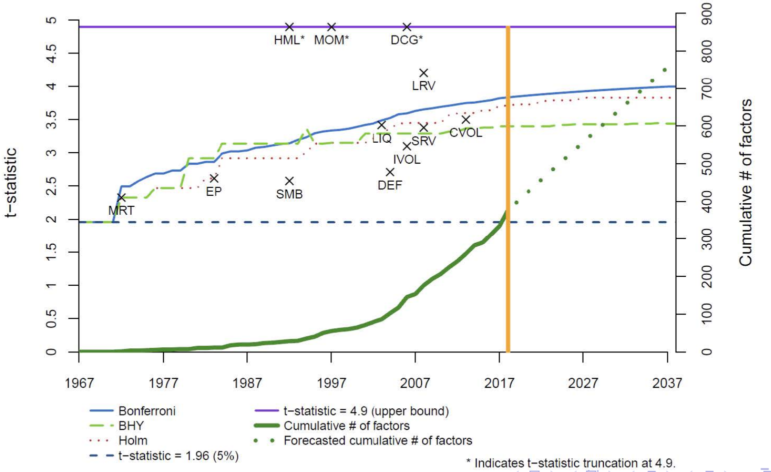
Test →	1	2	3	4	5	6	7	8	9	10	
t-statistic	1.99	2.63	2.21	3.43	2.17	2.64	4.56	5.34	2.75	2.49	3
p-value (%)	4.66	0.85	2.71	**0.05**	3.00	0.84	**0.00**	**0.00**	0.60	1.28	

Panel C: Holm adjusted *p*-values and "significant" factors

Reordered tests b	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	
Old order	8	7	4	9	6	2	10	3	5	1	4
p-value (%)	**0.00**	**0.00**	**0.05**	**0.60**	0.84	0.85	1.28	2.71	3.00	4.66	
$\alpha_w/(M+1-b)$ $\alpha_w=5\%$	0.50	0.56	0.63	0.71	0.83	1.00	1.25	1.67	2.50	5.00	

Factor “Zoo”

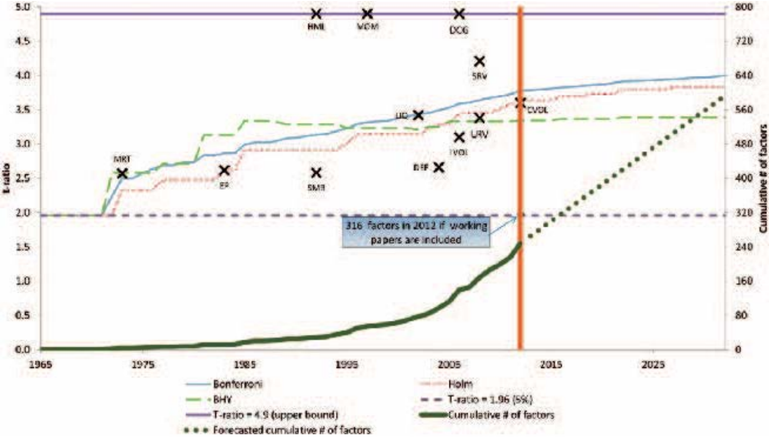
- **The Bonferroni Correction:** The blue dashed line is the two sigma (t-statistic = 2.0), which is common for establishing significance for a single test. As more factors were discovered, the threshold increases.



* Indicates t−statistic truncation at 4.9.

Factor "Zoo"

- Both Bonferroni and Holm adjusted benchmark t-statistics are monotonically increasing in the number of discoveries (Harvey, Zhang, and Liu, 2016). For Bonferroni, the benchmark t-statistic starts at 1.96 and increases to 4.00 in 2032.



Model Construction Methodologies for a Factor-Based Trading Strategy

- The key aspect of building a model is to (1) determine what factors to use out of the universe of factors that we have, and (2) how to weight them.

- We describe four methodologies to combine and weight factors to build a model for a trading strategy. These methodologies are used to translate the empirical work on factors into a working model.

- It is important to be careful how each methodology is implemented. In particular, it is critical to balance the iterative process of finding a robust model with good forecasting ability versus finding a model that is a result of data mining.

The Data Driven Approach

- A *data driven approach* uses statistical methods to select and weight factors in a forecasting model. This approach uses returns as independent variables and factors as the dependent variables.
- There are a variety of estimation procedures, such as neural nets, classification trees, and principal components, that can be used to estimate these models.
- Many data driven approaches have no structural assumptions on potential relationships the statistical method finds. Therefore, it is sometimes difficult to understand or even explain the relationship among the dependent variables used in the model.

The Factor Model Approach

- The goal of the factor model is to develop a parsimonious model that forecast returns accurately. One approach is for the researcher to predetermine the variables to be used in the factor model based on economic intuition. The model is estimated and then the estimated coefficients are used to produce the forecasts.

- A second approach is to use statistical tools for model selection. In this approach we construct several models - often by varying the factors and the number of factors used - and have them compete against each other, and then choose the best performing model.

- Factor model performance can be evaluated in three ways. We can evaluate the fit, forecast ability, and economic significance of the model.

The Heuristic Approach

- Heuristics are based on common sense, intuition, and market insight and are not formal statistical or mathematical techniques designed to meet a given set of requirements. The researcher decides the factors to use, creates rules in order to evaluate the factors, and chooses how to combine the factors and implement the model.

- There are different approaches to evaluate a heuristic approach. Statistical analysis can be used to estimate the probability of incorrect outcomes. Another approach is to evaluate economic significance.

- There is no theory that can provide guidance when making modeling choices in the heuristic approach. Consequently, the researcher has to be careful not to fall into the data mining trap.

The Optimization Approach

- In this approach, we use optimization to select and weight factors in a forecasting model. An optimization approach allows us flexibility in calibrating the model and simultaneously optimize an objective function specifying a desirable investment criteria.
- There is substantial overlap between optimization use in forecast modeling and portfolio construction. The factors provide a lower dimensional representation of the complete universe of the stocks considered.
- Besides the dimensionality reduction, which reduces computational time, the resulting optimization problem is typically more robust to changes in the inputs.

Machine Learning Approach

- The greatest endeavor in Asset Pricing has been documenting the properties of the stochastic discount factor (SDF) or pricing kernel (PK), which allows to price any asset with unknown future payoff, with the goal of understanding the determinants of asset returns.

- Multifactor models like Equation (3) raise also misspecification concerns. First, factor exposures are usually not allowed to change over time, that is, $\beta_{i,t} = \beta_i$ like in Fama and French (2015). Loadings might instead depend on macroeconomic (Ferson & Harvey, 1991) or asset-specific variables (Kelly et al., 2019).

- Second, the PK functional form is likely complex and unknown (Chen et al., Forthcoming): leading theoretical contributions postulate nonlinearities between returns and state variables (e.g., Campbell & Cochrane, 1999; Bansal & Yaron, 2004), which impact the return factor structure, too.

Machine Learning Approach

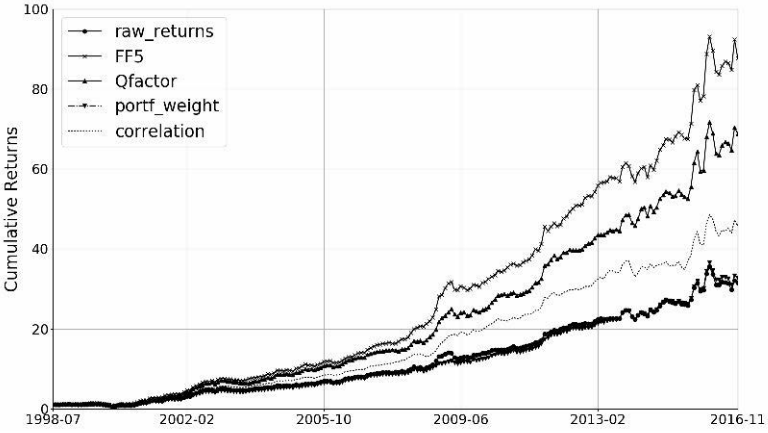
- Gu et al. (2020) provide a brilliant definition summarizing the main features of these techniques: they are generally used for high-dimensional predictions; they can be regularized to mitigate overfitting; their algorithms efficiently search and select among many model specifications.

  $$r_{i,t+1} = g(z_{i,t};\theta) + \epsilon_{i,t+1} \tag{8}$$

  The simplest formulation of Equation (8) assumes that returns are linear in characteristics: $g(z_{i,t};\theta) = z_{i,t}\theta$.
- To get the "big picture" sacrificing some details to prioritize the strongest predictors, reducing the number of parameters is vital. This can be done through regularization (or penalization), a key concept in ML and "one of the first signs of the existence of intelligent inference" (Vapnik, 1998).

Machine Learning Approach

- Sak et al. (2021) showed that the ML portfolio outperforms entrenched factor models, presenting a novel approach to understanding financial anomalies.



Python Sample - Fundamental Factor Long Short Strategy

- In this tutorial we implemented a long/short equity strategy based on fundamental factors. The idea comes from AQR white book: A New Core Equity Paradigm.
- The original version is a long only strategy. We developed it into a long/short version. The paper strategy used some fundamental data as measures of value, quality and momentum, and then ranked all the stocks in the universe according to the factors.
- The strategy only long the stocks ranking at the top, but our algorithm would at the same time short the stocks ranking at the bottom. This strategy consistently beats the market and has solid economic intuition.

  See Python Code.