WOLFE RESEARCH

LUO'S QES RESEARCH
Quantitative Research
Economics
Portfolio Strategy

MULTI-DIMENSIONAL ALPHA

February 8, 2017

THE BIG AND THE SMALL SIDES OF BIG DATA

QES Handbook of Active Investing, Part I

- **A Paradigm Shift in Active Investing.** Traditionally, investment management is divided into active and passive, discretionary (i.e., fundamental) and systematic (i.e., quantitative), top-down macro and bottom-up securities selection. The Big Data evolution gives us the access to vast amount of unconventional information – connecting people (e.g., management, board members, and analysts), products, financials, market data, and global economy – via structured, textual, imagery, audio/video, and location data. The incredible success of machine learning algorithms in fields from the GO game, virtual reality, to driverless cars has raised both hopes and fears among investment managers. The next frontier of active investing is far beyond corporate access and value/momentum factors that we are accustomed to. The next generation of active investing is about how to best integrate data and technology in your investment process and we are here to help.
- **A Research Series on Big Data and Machine Learning.** We are introducing our research in a sequence of four papers. We start with an introduction of data, technology, and data science. In the coming weeks, we will address signal research and multifactor models (from data to knowledge); machine learning, style rotation, and our global stock selection models (beyond the convention); and risk, portfolio construction, trade execution, and performance attribution (practical implementation).
- **The Big and the Small Sides of Big Data.** In this research, we first review the 30-year history of quantitative investing and highlight why changes are inevitable. We discuss the technology infrastructure that supports the Big Data transformation and then focus on the data contents. Lastly, we introduce data science, i.e., how to clean, process, and transform data into knowledge and actionable insights. We provide detailed coverage on advanced topics such as missing value imputation, currency and split adjustments, and building factors on a global universe.

Source: Yin Luo

Yin Luo, CFA, CPA
YLuo@wolferesearch.com

Javed Jussa
JJussa@wolferesearch.com

Sheng Wang
SWang@wolferesearch.com

QES Desk Phone: 1.646.582.9230
Luo.QES@wolferesearch.com

Table of Contents

A LETTER TO OUR READERS

Welcome to the first edition of the QES Handbook series

This research paper marks the official launch of our QES (Quantitative, Economics, and Strategy) research. In a sequence of four papers, we will outline our technology infrastructure and investment philosophy of active global equity investing in the new age. In a separate research series, we will discuss our framework for global macro, economics, and portfolio strategy research.

A PARADIGM SHIFT IN ACTIVE INVESTING

Traditionally, investment managers are in silos, divided into active and passive, discretionary (also called fundamental) and systematic (also known as quantitative), and top-down global macro and bottom-up securities selection. Managers and research analysts are structured by countries, sectors, and styles. The current structure ensures efficiency, but prevents us from information sharing.

Currently, fundamental managers rely on in-depth valuation analysis of subject companies (primarily based on financial statement data), interviewing company management, and discussing with industry experts to form a view on the trend of in the industry and firms they want to invest. Similarly, quantitative managers build multi-factor models exploring market anomalies and take advantage of the breadth and diversification benefit.

The Big Data evolution, however, has changed the playing field dramatically. We can now link the subject company with its customers, suppliers, competitors, joint ventures, and other partners. We can track and analyze key personnel (C-suite, board members and other insiders, sell-side analysts, and institutional shareholders and creditors) and connect them together. We can trace the products and services each company provides and then link to the changes in demographics and consumer spending patterns. Satellite imagery and mobile location tracking allow us to pinpoint the exact activities of a company, a shopping mall, an industrial site, an oil field, or a country. Even the traditional fundamental data now goes far beyond the three sets of financial statements – we can drill down to the specialized industries (e.g., banks, insurers, utilities companies) or countries (e.g., China, Japan). Unstructured data such as textual information (newspapers, websites, blogs, research reports, academic papers) and audio/video files are being presented to us in a way that was unimaginable even just a few years ago. How to analyze this mountain of data and form a consistent investment opinion poses a huge challenge to portfolio managers. We hope to shed some light on how to manage and utilize Big Data in this paper.

Furthermore, artificial intelligence, machine learning, and computer algorithms have made enormous success in fields like medical research, fraud detection, virtual reality, and driverless cars. It has raised both hopes and fears among portfolio managers. Does technology provide us with extraordinary tools to identify market anomalies, or are computer algorithms replacing human analysts and traders? Big Data and machine learning present new challenges and opportunities. In this series of papers, we plan to introduce this important new paradigm shift of investment philosophy to our clients. Our long-time readers should know our style – we are not interested in the big high level discussion, technical jargons, and cliché words – rather, we will focus on the practical issues of how to implement these new models and techniques in our investment process.

This paper focuses on the first basic building block of active investing – data and data science. The next paper will address signal research and multi-factor models, i.e., from data to knowledge. Then we will move on to the more advanced topics – from machine learning, style rotation to sophisticated alpha models. Lastly, we will elaborate on the practical issues of portfolio construction, trade execution, performance attribution and hedging.

Data is the foundation of investing. Traditionally, equity portfolio managers follow either a quantitative or a fundamental approach. Despite the many efforts of bringing them together at many firms, it remains a significant challenge, due to the very different investment philosophy, process, and operations. In recent years, the intensified competition, the threat from passive investment/ETF, and the rapid development in Big Data and machine learning techniques have changed the investment world in a profound way. We are seeing more and more quantitative managers who are willing to dive deeper and incorporating fundamental views into their portfolio, e.g., via alpha capture or quantifying fundamental views. Similarly, an increasing number of fundamental managers are embracing the Big Data evolution and want to use quantitative models to improve their investment process.

HOW CLIENTS CAN USE OUR RESEARCH

For fundamental managers, generally speaking, making new investments in data, technology, and quantitative modeling is a significant undertaking. Not only does it take considerable time, budget, and efforts, it is also difficult to compete with established quantitative firms. We provide the access of comprehensive data contents, research ideas and data feeds to support your investment process.

For systematic managers, the competition for new data, new ideas and new modeling techniques require tremendous investment in data subscription, R&D, and portfolio implementation. Our research and services not only help you with access to the most cutting edge ideas and data, but also can lead to substantial savings in time and cost.

PUBLICATION SERIES

We brand our research as QES, representing three areas in Quantitative Research, Economics, and Portfolio Strategy. Our research comprises four publication series:

- ***Multi-Dimensional Alpha***. We plan to publish 10-15 in-depth research papers per year, with detailed discussion on modeling techniques and research methodology, on a wide range of topics from Big Data and factors, sophisticated machine learning techniques, to risk and portfolio construction, across all asset classes.
- ***Current Affairs***. This monthly/weekly series focuses on timely and actionable ideas on global stock selection and global macro.
- ***Journal of QES***. This is a monthly publication that highlights the latest and most relevant academic research papers, sourced from both working paper repositories and peer-reviewed journals. Please see Luo, et al [2016] for a recent example.
- ***Luo's QES Newsletter***. This is a monthly newsletter on the latest development in the quantitative and macro investing world.

ASSET CLASSES

Our research covers three areas in quantitative, economics, and portfolio strategy in all asset classes. The main focus is on global equities – global all capitalization for both developed and emerging markets. In addition to stock selection, we will also have comprehensive coverage on global economic forecasting, GTAA (Global Tactical Asset Allocation), G10 currency, commodities, country and sector allocation, and style/factor rotation.

FORTHCOMING RESEARCH

Due to size limit, this paper only covers the first part of systematic investing. In the next few weeks, we will publish the other three key components:

- *Signal Research and Multifactor Models* – From Data to Knowledge
- *Machine Learning, Style Rotation, and the Next Frontier in Systematic Investing*
- *Risk, Portfolio Construction, Trade Execution, and Performance Attribution* – From Theory to Practice

In addition to the four-part introduction of Big Data and Machine learning in global equity investing, we are also working on a number of issues:

- From Nowcasting to Forecasting – Economics and Portfolio Strategy in the New Age
- Industry-Specific Models in Global Banking and Insurance Industries
- Accounting Quality, Fraud Detection, and Corporate Governance
- Factors based on Alternative Data Sources
- Machine Learning in Global Stock Selection

DATA FEEDS AND CONSULTING SERVICES

After building our technology infrastructure on our own, we understand the tremendous time and efforts required. Not all clients have the resources to construct their own infrastructure. Even for clients who build and maintain their own systems, it might still be much more cost effective to access our data than building on their own. Subscribing to data, building the system, maintaining your database, searching for signals all take time. We plan to offer custom data feeds, based on many of our research publications. Please contact us for details.

We look forward to working with you all at our new home. Any feedback and suggestion are more than welcome!

Regards,

Yin, Javed, Sheng, Kartik, and Luo's QES team

THE PAST AND THE FUTURE OF SYSTEMATIC INVESTING

In this section, we briefly review the past 30-year history of quantitative investing. We then discuss the challenges ahead of us. The availability of abundant data across a wide range of areas in vastly different structures provides us both opportunities and threats. Active managers face intense competitions from not only other managers, but also passive and smart beta indices. The ability of machine learning algorithms seems to be limitless, giving us both aspiration and despair.

THE HISTORY – 30 YEARS OF QUANTITATIVE INVESTING

In the past 30 years, the systematic investing industry has gone through significant changes. We can roughly split the history into three periods.

The Early Years in 1980s and 1990s

Arguably, the first real quantitative investment fund is Wells Fargo's dividend tilt fund started in 1978. However, quantitative investing was on the sidelines until early 2000s. The burst of the technology bubble caught many active investors off guard. The disappointment towards traditional stock picking, the irrational exuberant nature of human behavioral biases in investing, along with the availability of company fundamental databases and computing power triggered the start of the "Golden Years" of quantitative investing from early 2000 until the summer of 2007.

The Golden Years of 2000-2007

During the golden years, the performance of most quant funds was extremely strong, especially after adjusting for risk. The stellar performance attracted tremendous assets. With hindsight, the models used at the time were relatively simple – a mix of value and momentum, earnings revisions, and cash flow based signals. More problematically, most quantitative managers had similar factors, models, portfolio construction techniques, and trade in similar fashions, which exposed the industry to potential crowded trades.

Summer 2007 Quant Crisis and the Subsequent Risk-on/Risk-off Environment

Indeed, in the week of August 1 to August 10, 2007, value and momentum[1] plunged -7% and -4% in less than two weeks, respectively (see Figure 1 A). At the time, it was considered as a fairly dramatic drawdown in such a short period of time. Ironically, both factors recouped the loss in the next two weeks. It is now generally accepted[2] that the considerable swing of quant factors in the summer of 2007 was caused by a sudden liquidation from a few multi-strategy funds and proprietary trading desks, possibly due to margin calls or risk reductions in other positions outside of their quant equity books. The initial liquidation triggered losses at many quant funds, due to the similarity in their models, which exacerbated further sell-offs of stocks with the same characteristics.

While many quant investors were still struggling to comprehend the implications of the quant crisis and crowded trades, the subsequent 2008 financial crisis had changed the landscape once and for

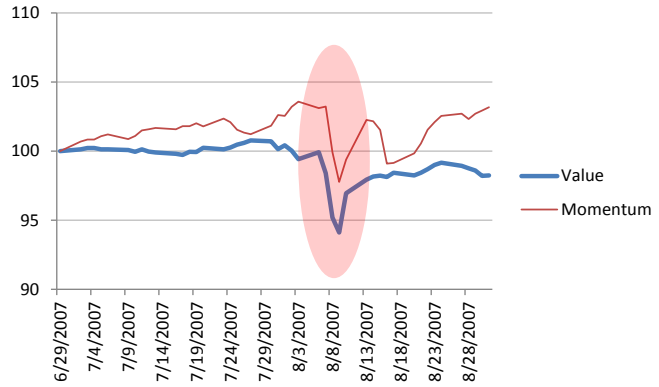
[1] We use trailing earnings yield (trailing 12-month EPS/price) and 12-month total return excluding the most recent month to represent value and momentum, respectively. Value and momentum factors are both constructed as simple long/short quintile portfolios, where we buy the top 20% of stocks with the cheapest valuation (or highest price momentum) and short the bottom 20% worst stocks, equally weighting stocks in both long and short baskets. Portfolios are rebalanced monthly and transaction costs are not included.

[2] See Khandani and Lo [2007] for detailed discussion of the summer 2007 quant crisis. We have done extensive research on strategy crowding, using a wide range of metrics from short interest to trading patterns (see Cahan and Luo [2013] for one example).

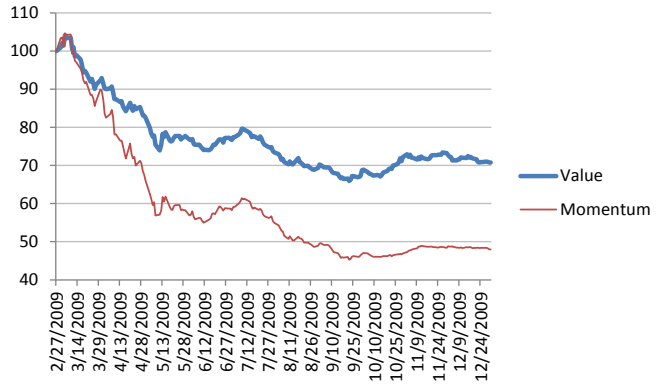
all. Quant funds initially benefited considerably from the onset of the 2008 global financial crisis, with the general nature of overweighting stocks with higher quality, lower risk, and cheaper valuation. The US equity market reached a bottom on March 9, 2009. Then, while the economy still struggled during the recession, the market sentiment turned the other way swiftly, as investors embraced for a quick economic recovery. As a result, risky stocks rallied, while low risk assets massively underperformed. Quant factors, in particular, momentum and low risk plunged in the March-May 2009 risk rally (see Figure 1 B). In less than three months from March 9, 2009 to June 1, 2009, the price momentum factor suffered a loss of over -45% and similarly, the low beta strategy went down over -50%. The loss was so severe and completely overshadowed the summer 2007 quant crisis.

Figure 1 Two Episodes of Quant Crisis

A) Summer 2007



B) March-May 2009 Risk Rally



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

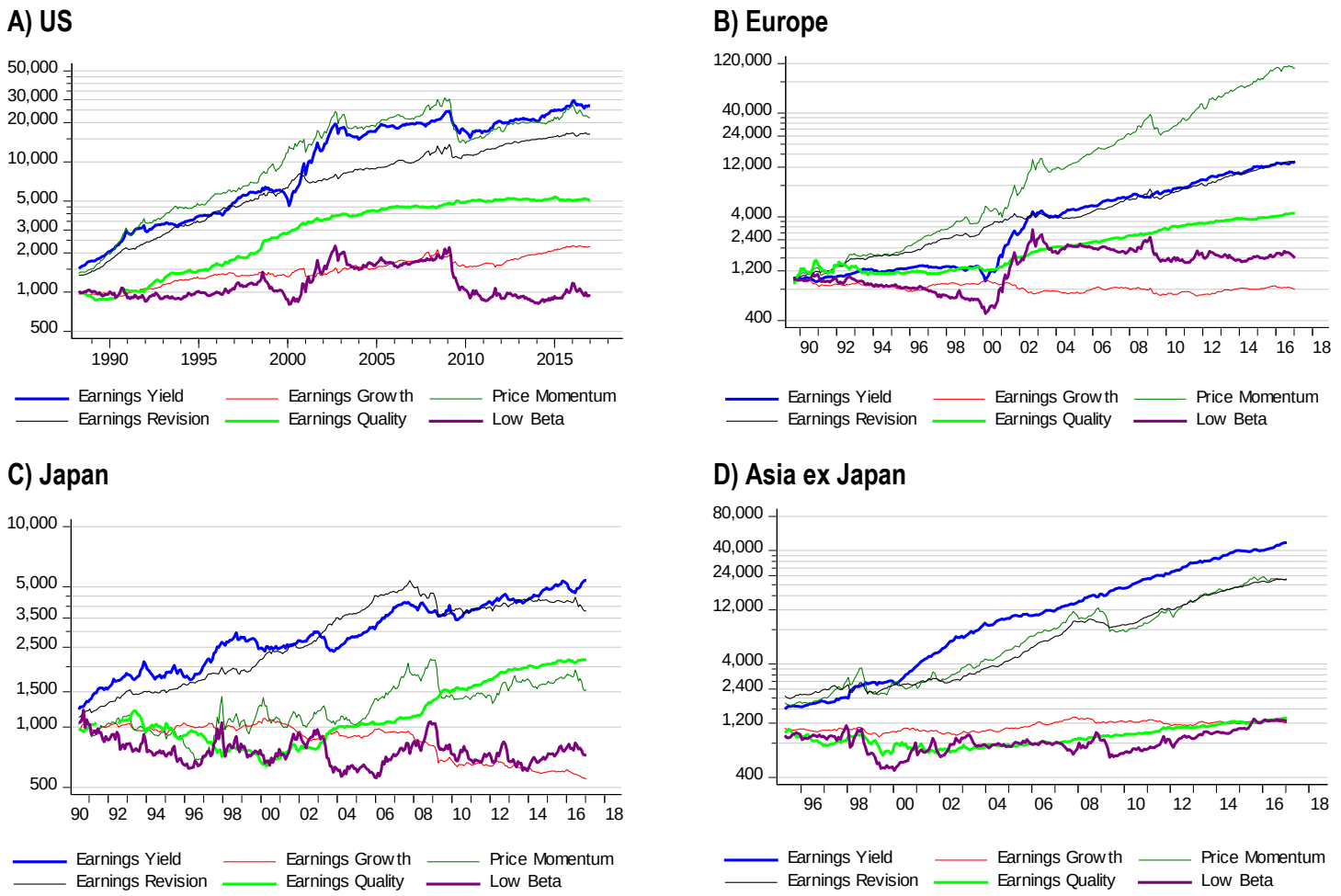
Beyond the two well-known quant crises, as shown in Figure 2, the performance[3] of six common stock selection factors in four major regions of the world (US, Europe, Japan, and Asia ex Japan) reveals a few interesting patterns:

- Factor performance was much stronger in the early years prior to 2007. Post 2008, in the US and Japan in particular, returns compressed, while risk exacerbated considerably.
- Most factors deliver positive returns in the long term, but they are subject to periodic drawdowns. The downside risk can be substantial, especially for price momentum and low beta strategies. The low risk factor is the most volatile factor in almost all regions.
- It tends to be far easier to generate alpha in Europe and AxJ than in the US and Japan.

[3] We will define these factors and our backtesting methodology in more details in a forthcoming research, *Signal Research and Multifactor Models*. On a high level, we divide the world into nine regions (US, Canada, Europe, UK, Asia ex Japan, Japan, Australia and New Zealand, LATAM, and emerging EMEA. We include both large- and small-cap stocks in our investment universe. Factor performance is measured using long/short quintile portfolios, where we long the best ranked stocks (equally weighted) and short the worst ranked stocks (also equally weighted). The portfolios are constructed on a country/sector neutral basis. Portfolios are rebalanced monthly without taking into account of transaction costs and short availability. We use local currency to compute factors and stock returns.

- In the US, value has the highest cumulative return, while accounting quality has the best risk-adjusted performance. In Europe, price momentum factor dominates in terms of returns, while accounting quality has the highest Sharpe ratio. In Japan, price momentum, earnings growth, and low beta anomalies virtually do not exist, while value and earnings revision have reasonable performance. In AxJ, the classic value, price momentum, and earnings revision have all produced decent performance.

Figure 2 The Performance of Common Stock-Selection Factors

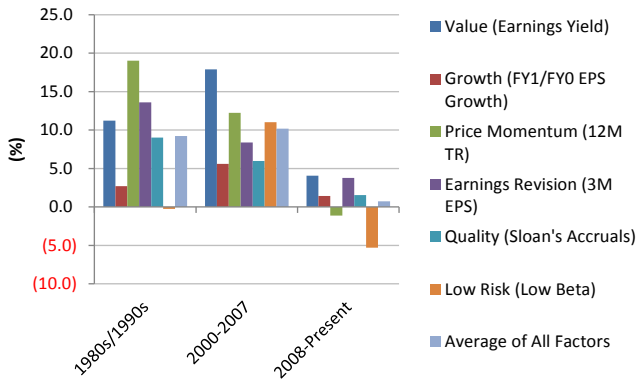


Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

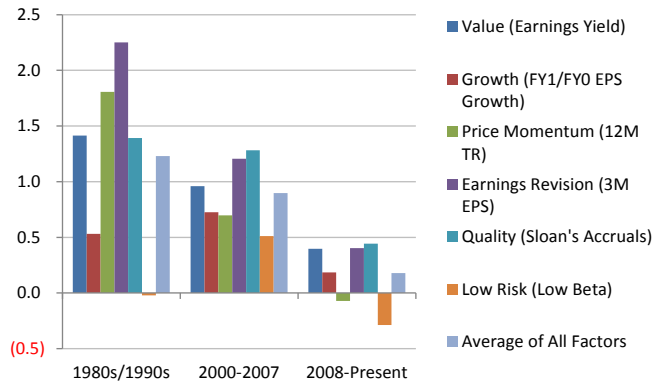
In the US equity market, the performance of most common factors (see Figure 3 A and B) clearly shows significant decay in the post 2008 period. Globally, in all major regions, factor performance has declined in recent years, especially in the US and Japan (see Figure 3 C and D).

Figure 3 The Challenges Ahead of Us

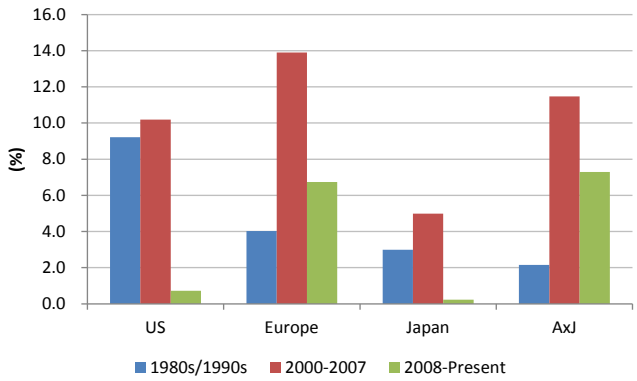
A) Average Factor Return in the US



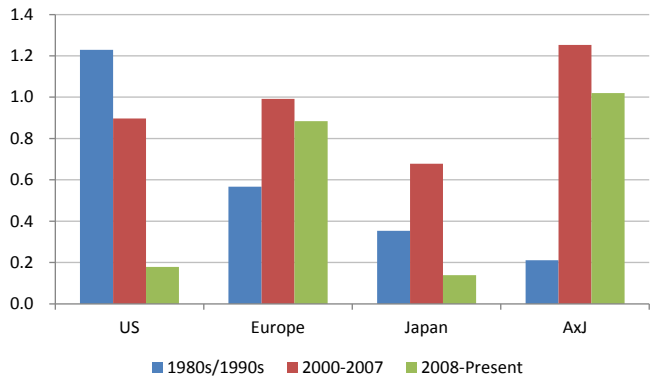
B) Average Sharpe Ratio in the US



C) Average Factor Return Globally



D) Average Sharpe Ratio Globally



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

THE CULPRIT AND THE VERDICT

It has long been debated whether the decline of factor performance in recent years is transitory or permanent. We have always argued that the challenge is secular and the good old days are over, for a number of reasons:

- Intensified competition and market efficiency are reflected in rising correlation among factors in the same region and among regions for the same factor, coupled with heighted downside risk.
- Geopolitical risk and uncertainty are playing an increasingly dominant role in investing. The investment world is marked by periodic risk-on and risk-off, which dictates the performance of investment styles (and factors).
- Factor payoff patterns are becoming perceptibly nonlinear. The traditional Fama-French [1993, 1996] type of linear factor models no longer captures the cross sectional risk and return tradeoffs very well.
- *Ad hoc* factor selection and static factor weighting are likely to face the biggest headwind.

Rising Factor Correlation and Heightened Downside Risk

As shown in Figure 4 (A), the pairwise correlation among factors (in the same region) has either increased significantly (in Japan and AxJ) or remains high (in the US and Europe) in recent years. Furthermore, the correlation of the same factors across regions has spiked even higher (see Figure 4 B). The surge of cross-regional correlation for the price momentum and low risk factors is particularly noticeable. The spillover of financial and political risk across regions is at a speed that we have never seen before. The conventional wisdom of diversifying across factors and regions is being heavily scrutinized.

Much of the struggle, especially for the US market, however, is due to the risk rally in March-May 2009. In less than three months from March 9, 2009 to June 1, 2009, the price momentum and low beta strategy went down almost -50%[4]. As shown in Figure 4 (C) and (D), the downside risk of most factors in the US has extended in recent years and remains high in Europe and Japan.

[4] Price momentum essentially invests in stocks with highest past 12-month returns (and shorts the ones with the lowest returns). At the bottom of the market on March 9, 2009, stocks with the best performance tended to be mostly low beta stocks (in a bear market environment). Therefore, price momentum and low beta was essentially the same factor at the time. Indeed, the correlation between the two factors from March 9, 2009 to June 1, 2009 was around 99%.

Figure 4 Rising Factor Correlation and Heightened Downside Risk

Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

The Impact of Risk-on/Risk-off

The post-2008 period is marked by a periodic shift of investor's risk sentiment. A common phrase used in the investment industry is risk-on/risk-off[5]. There is not a unanimous definition, but generally speaking, risk-on refers to an environment that investors are optimistic about the future and therefore are willing to invest in risky assets. On the other hand, in a risk-off regime, investors worry about the underlying investment environment and stay away from risky stocks.

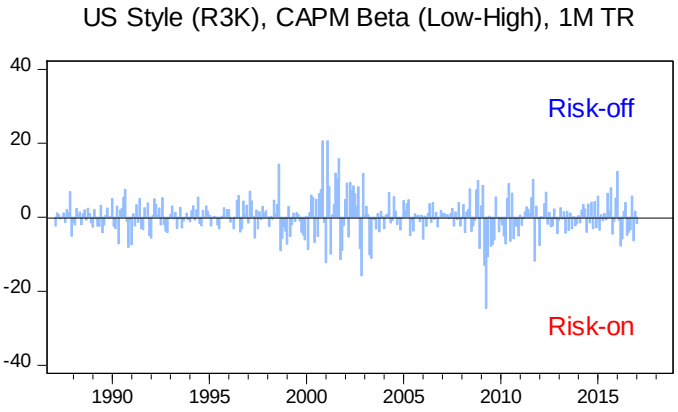
A simple way to capture risk-on/risk-off is to examine the return of the low beta factor. By construction, the low beta factor invests in the top 20% of stocks with the lowest beta and simultaneously shorts the bottom 20% of stocks with the highest risk. When the return of the low beta portfolio is positive, it indicates a risk-off regime (i.e., investors pile into low risk stocks), and vice versa for positive return periods as risk-on.

[5] The phrase "risk-on and risk-off" has become such a cliché in recent years. For lack of better names, we shall continue to use it in this paper.

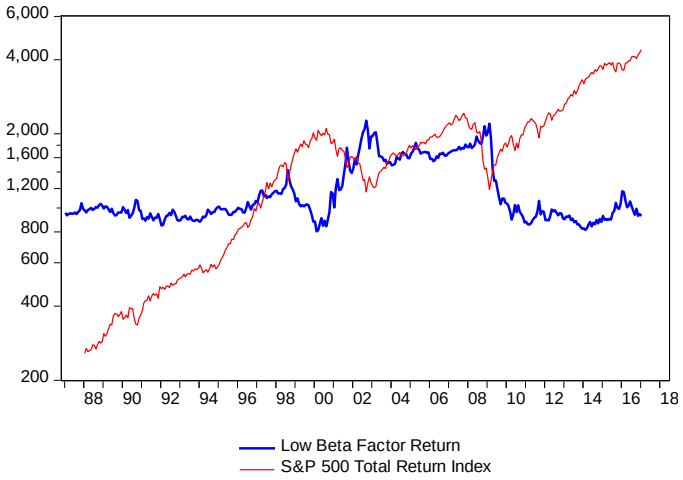
As shown in Figure 5 (A), it is evident that our simple risk-on/risk-off indicator captures the major market turning points – consistent positive returns (risk-off) during early 2000 (the burst of tech bubble) and 2008 (financial crisis) and similarly, consistent negative returns (risk-on) in late 1999 (tech bubble) and March-May 2009 risk rally. Although the low beta factor is negatively correlated[6] with the market (see Figure 5 B), they are different. The market and the low beta factor can rally (or dip) at the same time.

Figure 5 A Simple Risk-on/Risk-off Indicator

A) The Return of the Low Beta Factor



B) Low Beta Factor versus the Market



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

One problem of using our risk-on/risk-off factor directly is that the indicator is a little too noisy. To extract the true signal from noise, we apply a Markov Regime Switching (MRS) model and use the smoothed[7] probability as our improved risk-on/risk-off regime classification. As shown in Figure 6 (A), now the smoothed risk-on/risk-off regime is much more stable. Indeed, the estimated probability of remaining in a risk-on (risk-off) regime is 94% (97%). The average duration of risk-on (risk-off) regime is about 18 (38) months. The average return of the low beta factor is significantly lower at the risk-on regime (-3.1% per year) than at the risk-off environment (2.2%).

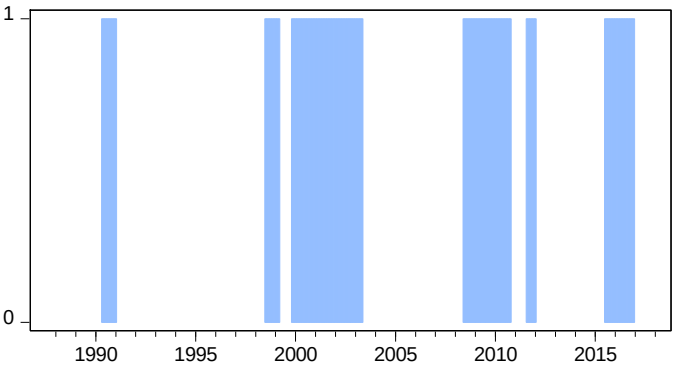
For most factors, the performance is weaker in a risk-on environment, but more importantly, the risk (i.e., dispersion in return distribution) and in particular, the downside risk tends to be materially higher in risk-on regimes. As shown in Figure 6 (B) and (C), the return distribution for both value and momentum factors is much flatter in risk-on regime, with a considerably longer left tail (i.e., negative return). The frequent risk-on/risk-off switches in the post-2008 period are the main reason behind the turbulent performance for many quant funds. Lastly, as shown in Figure 6 (D), volatility tends to be multiple times higher in risk-on regimes for almost all common factors.

[6] The correlation is about -69%.

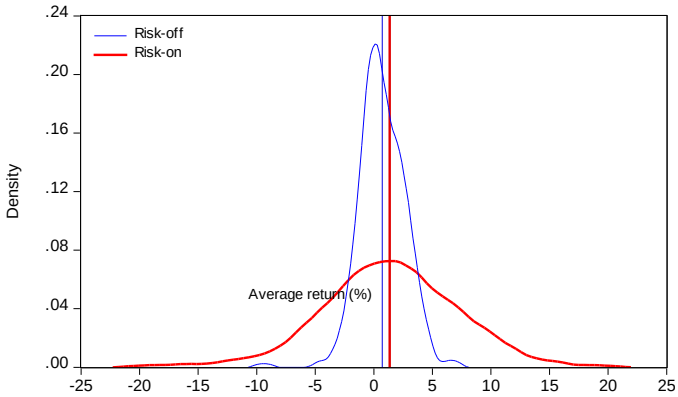
[7] Technically speaking, there are four regime probabilities: true out-of-sample, one-step-ahead forecast, filtered, and smoothed. The smoothed probability is estimated using the entire dataset; therefore, it provides the most precise estimate. However, the smoothed estimate is in-sample in nature and can't be used in real-time forecasting. For our purpose, as we try to understand the impact of risk-on/risk-off rather than making real-time prediction, it is better to use the smoothed probability.

Figure 6 Risk-on and Risk-off in the US Equity Market

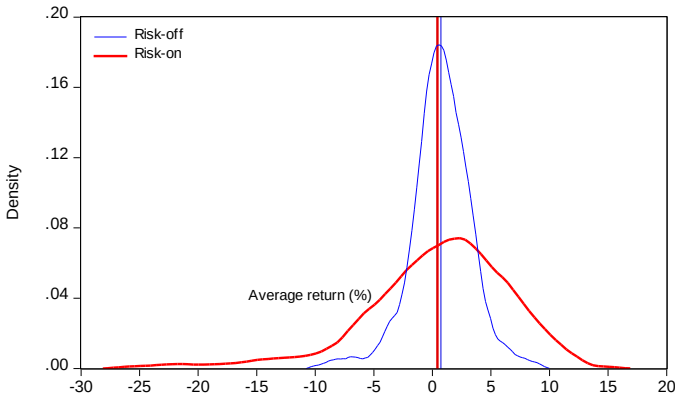
A) Risk-on/Risk-off Regimes (1=Risk-on)



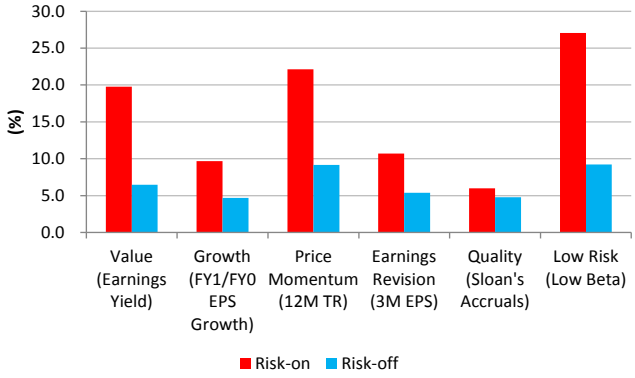
B) Performance Distribution – Value



C) Performance Distribution – Momentum



D) Factor Volatility in Risk-on and Risk-Off Regimes



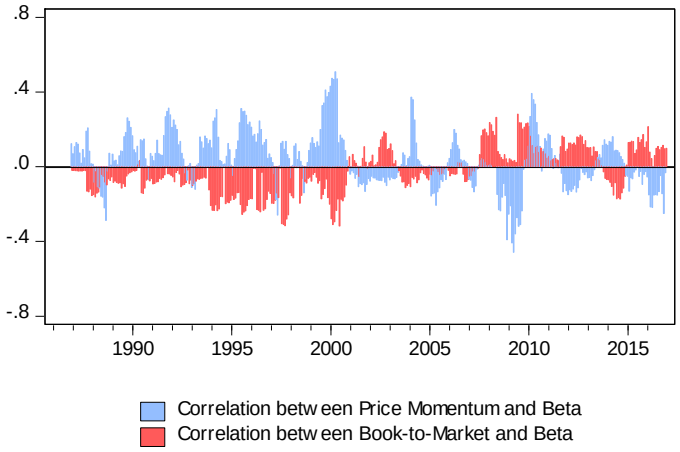
Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

In a risk-on/risk-off world, many factors are simply proxies for risk (either risk seeking or risk averse). For example, 2016 was an interesting year, as the performance of quant funds was very binary – either very strong or very poor – all depends whether they were on the right or wrong side of the risk regime switches.

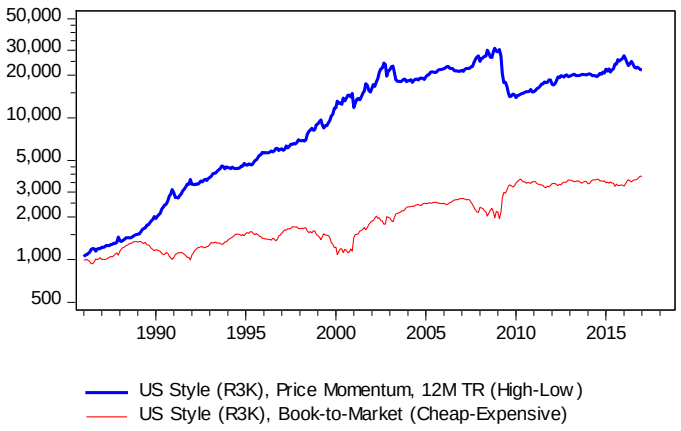
Two factors dominated the performance of most multi-factor models – value (proxied by book-to-market) and momentum (e.g., 12M return excluding the most recent month) in 2016. As shown in Figure 7 (A), book-to-market was positively correlated to beta, in a sense that value stocks were cyclical and benefited from an improvement in investors' risk appetite. On the other hand, price momentum was negatively exposed to beta – winners stocks were mostly low risk, which underperformed considerably since February 2016. The performance of book-to-market and price momentum forms a perfect mirror image (see Figure 7 B). Managers thought they invested in multi-factor models, but in the end, most portfolios resembled unidimensional market timing in 2016.

Figure 7 Cross-Sectional Correlation to Risk

A) The Correlation of Price Momentum and Book-to-Market versus Beta



B) The Cumulative Performance of Price Momentum and Book-to-Market



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

Factor Payoff Patterns are Becoming Increasingly Nonlinear

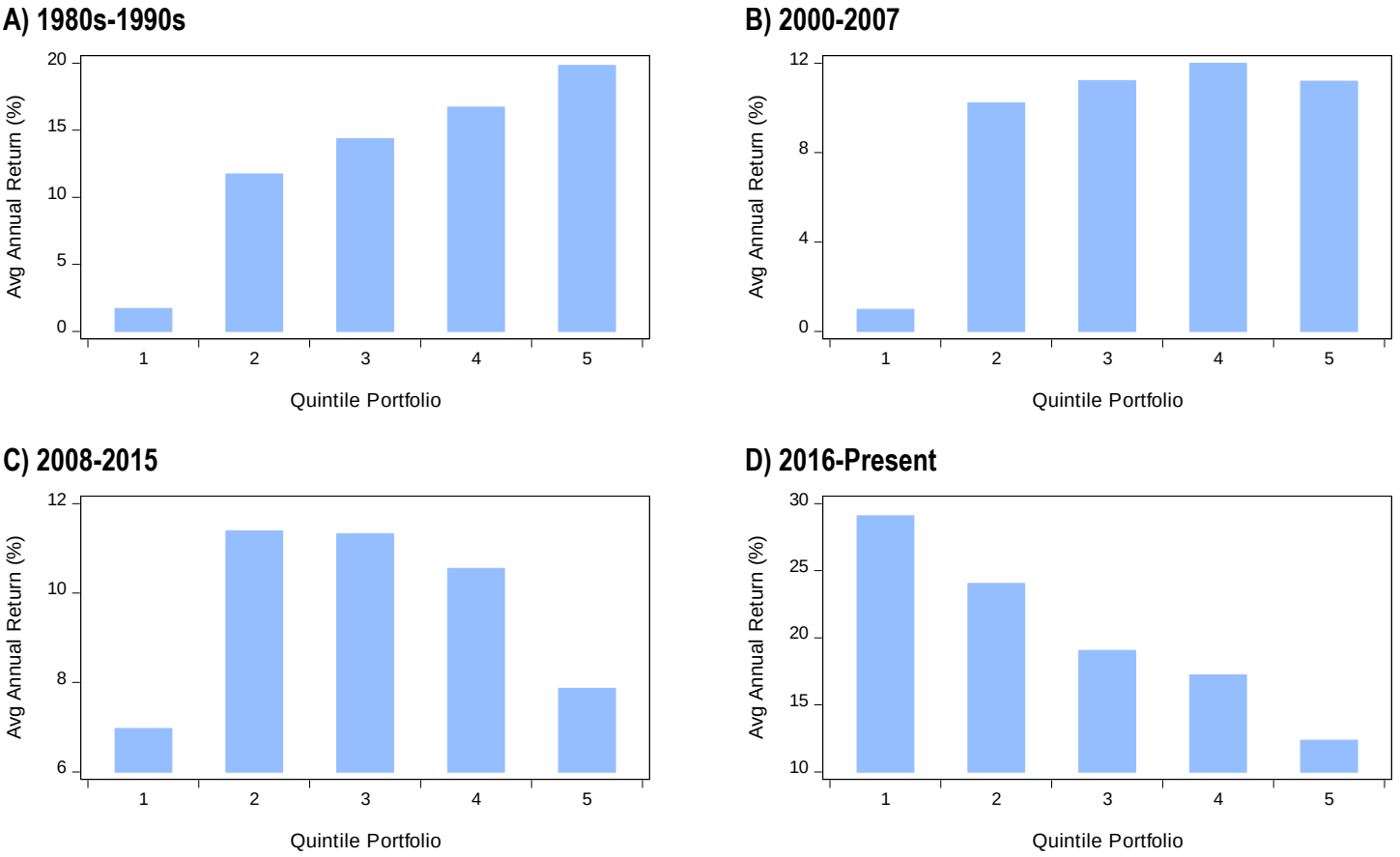
Many of the traditional market anomalies or stock-selection factors were discovered and the underlying academic papers were published before 2007. At the time, the relationship between factors and forward stock returns was primarily linear or at least monotonic. Not surprisingly, the predominant modeling techniques were also linear in nature, e.g., OLS regression, mean-variance optimization, etc.

As the market evolves, possibly due to a combination of arbitrage by investors and changes in the underlying market regimes, the payoff patterns are becoming progressively nonlinear.

Figure 8 (A) to (D) show the payoff patterns for one of the cornerstones of quantitative investing – price momentum, over four periods. We form five quintile portfolios, based on the month end price momentum factor. Then we rebalance the portfolio monthly. The four graphs illustrate the average returns of the five momentum quintile portfolios over time. If the payoff pattern of the price momentum factor conforms to the Jegadeesh and Titman [1993] study, we should expect a linear monotonic upward trend. In the early years in the 1980s-1990s (see Figure 8 A), that was exactly what we would expect, albeit Quintile 1 portfolio had a disproportionally low return, possibly due to limit arbitrage[8]. In the golden years of 2000-2007, the pattern became much less linear, but low momentum stocks in Quintile 1 still massively underperformed; therefore, investors who had shorted poor momentum stocks would have generated outsized returns. In the third period from 2008-2015, the pattern resembled an inverted U-shape, where both poor momentum stocks in Quintile 1 and best momentum firms in Quintile 5 underperformed the middle three quintile portfolios. In 2016, the pattern completely reversed to a monotonic downward trend.

[8] Shorting was more difficult and costly in the 1980s-1990s.

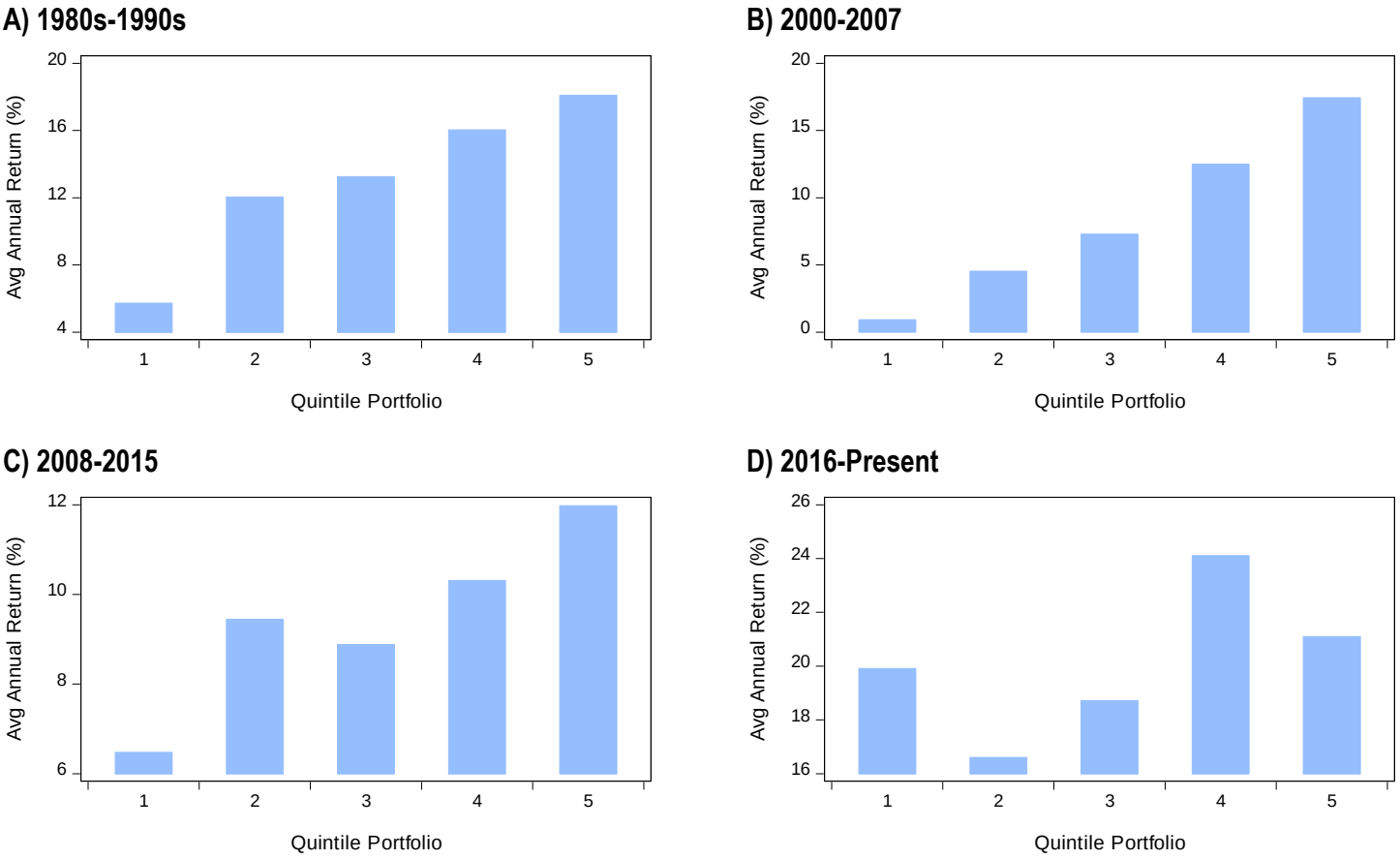
Figure 8 Price Momentum Factor in the US



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

The nonlinear pattern is not limited to price momentum. In fact, in the US market, the structure for most common factors has changed dramatically in recent years. Figure 9 (A) to (D) illustrate the changes of value factor performance in the US.

Figure 9 Value Factor in the US

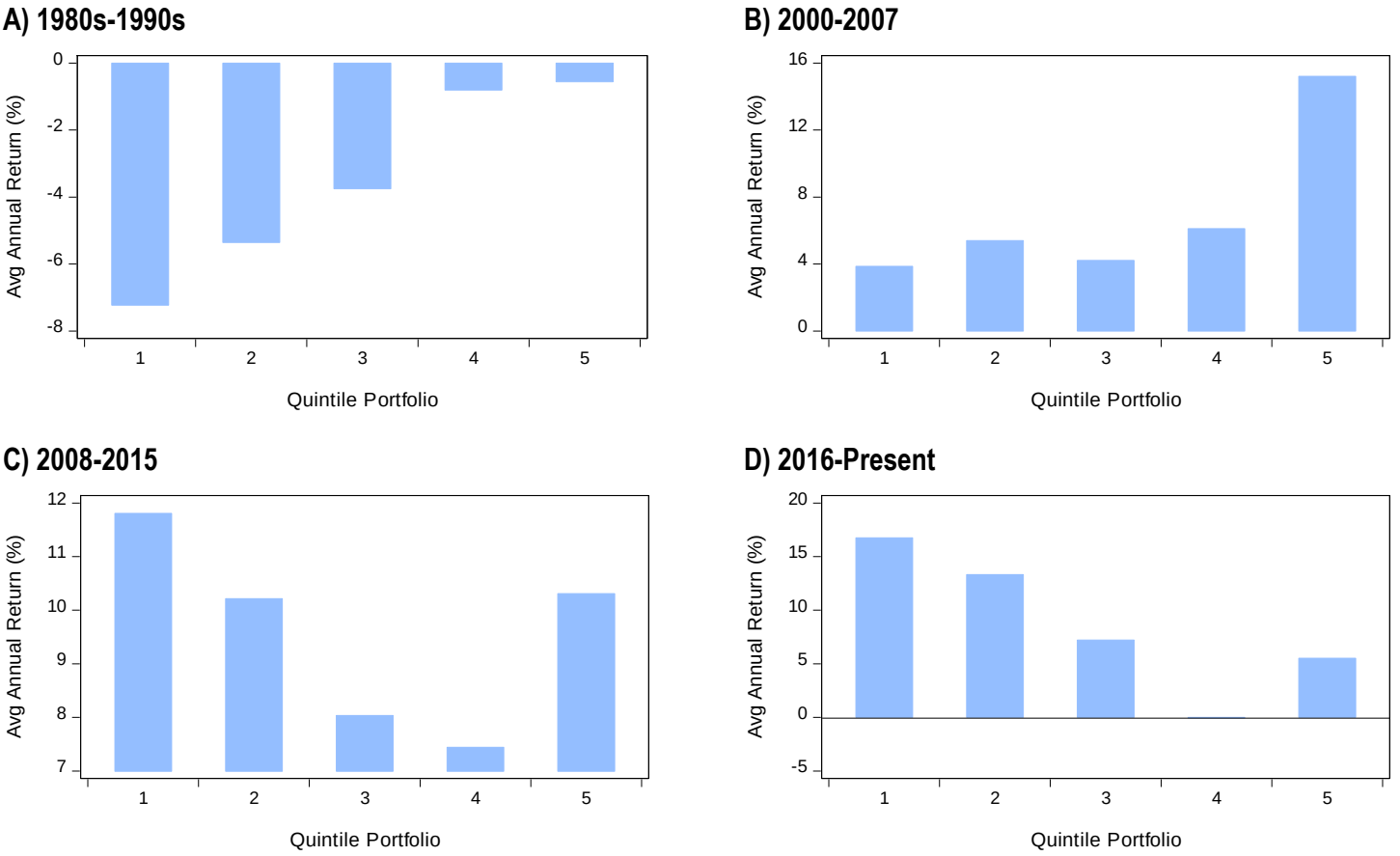


Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

Outside of the US equity market, it is generally not as severe as in the US. However, we also notice the patterns for many factors in many regions have changes, e.g., earnings revision factor in Japan (see Figure 10) and value factor in Europe (see Figure 11).

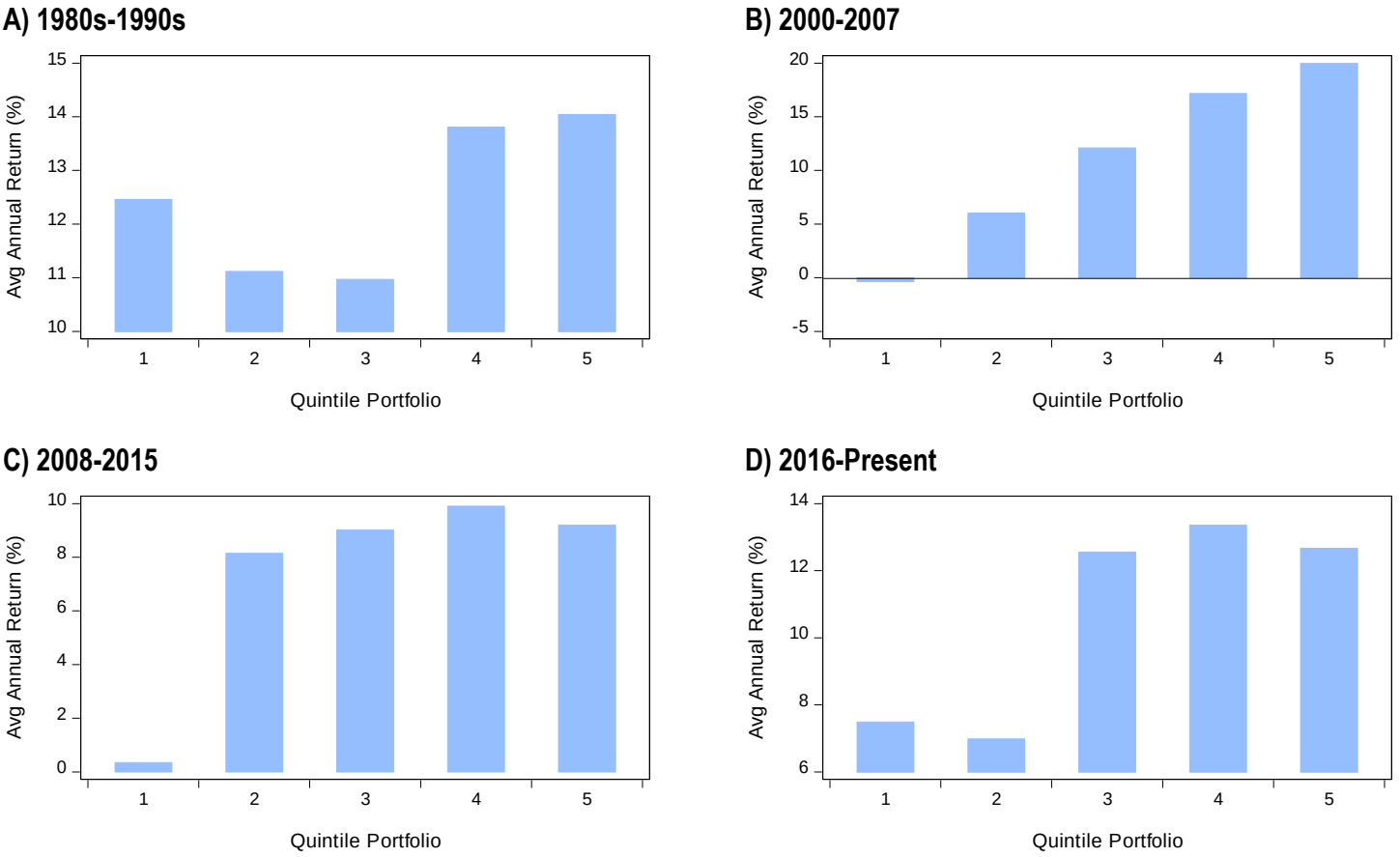
Extending our modeling techniques beyond linear regression poses a series of challenges to managers. It is not due to lack of nonlinear algorithms – in fact, there are far too many nonlinear models to choose from. Mainstream finance research is still predominantly linear in nature. Nonlinear models are often labeled as and confused with data mining. Even those limited research papers that reveal nonlinear patterns are *ad hoc* in many ways.

Figure 10 Earnings Revision Factor in Japan



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

Figure 11 Value Factor in Europe



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

Static Factor Models are Likely to Struggle

As a result of declining factor performance, changes in macro regimes, rising factor correlation, and the evolving nonlinear patterns, *ad hoc* factor selection and static factor models face particular difficulties in recent years.

To demonstrate, we conduct a simple simulation. We pick around 120 signals from our factor library[9] with complete history going back to 1986 in the US. For Europe, we have about 140 factors with history from 1992. We then backtest these factors in two separate periods: from 1986 to the end of 2007, and then post-2008.

As shown in Figure 12 (A), we do observe reasonably strong pattern of long-term factor momentum, i.e., those factors that had the strongest performance in 1986-2007 also tend to have the largest returns in 2008-present. However, there are two particularly worrying observations:

- The average performance of these 120 factors is 35% lower in the post 2008 era.

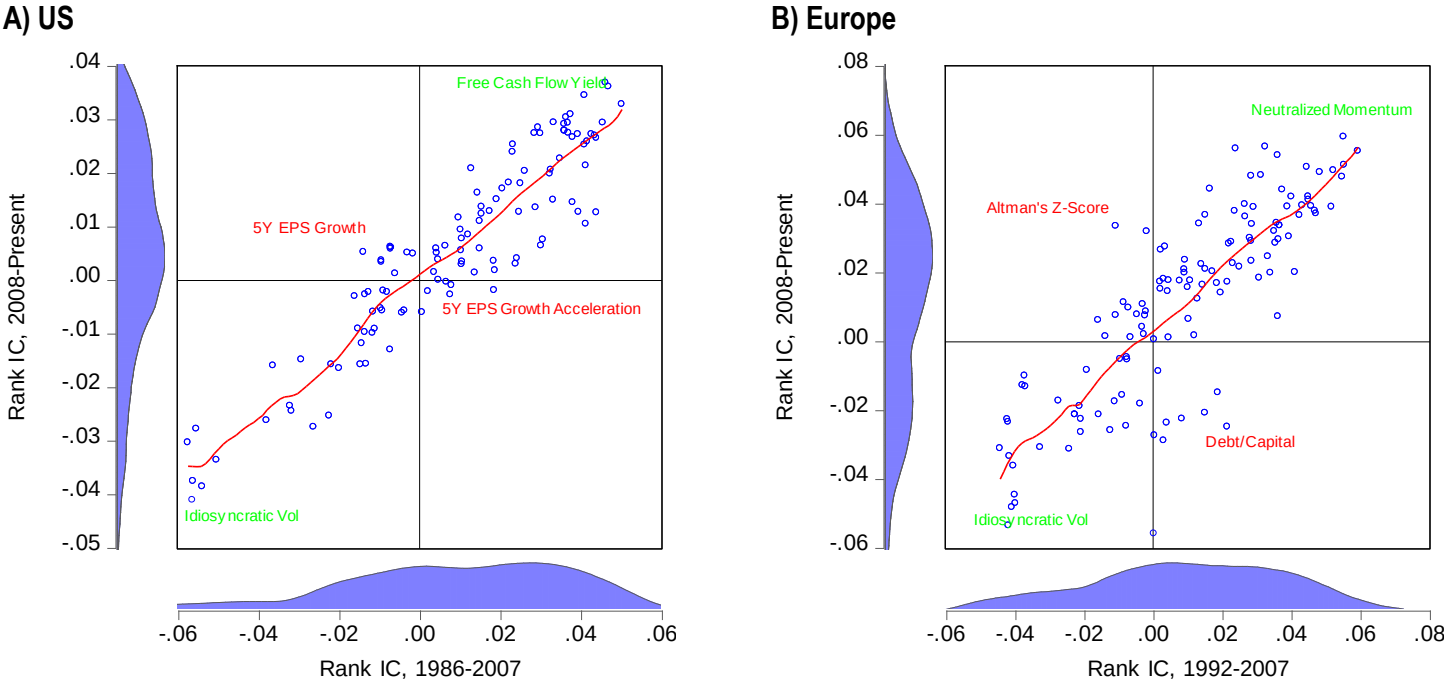
[9] We have about 400 relatively unique factors in our factor library.

- Almost 13% factors show different directions in the two phases. For example, historical five-year EPS growth factor in the 1986-2007 period was a negative signal, meaning companies with high growth underperform. However, in the 2008-present period, it becomes a positive signal, i.e., high growth stocks performing better.

In Europe (see Figure 12 B), we have different challenges:

- The average factor performance in the post 2008 period is actually 12% higher than it in the 1992-2007 era.
- However, 18% factors show different signs in the post 2008 period. The classic examples are financial leverage and default factors, e.g., debt/capitalization and Altman's z-score. In the early years (1992-2007), companies with higher leverage lead to higher stock returns, while these returns reverse the signs in the post 2008 period, as debt crisis looms in many parts of the continent.
- Factor momentum is much weaker in Europe than in the US, hinted by the looser distribution around the fitted line in the middle.

Figure 12 Factor Performance before and post 2008



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

The Latest Development

Both investment managers and asset owners have been thinking hard in the past few years about the future of active investing. Along the way, a couple of views have emerged:

- **Dynamic Factor Weighting and Style Rotation.** As we will elaborate in a follow-up paper, the factor selection and factor weighting decision at most quantitative managers was

predominantly *ad hoc* prior to 2008. In the post 2008 era, dynamic factor weighting has received great interest. However, factor timing is still mostly done in a discretionary way.

- **New and Uncorrelated Data.** Many managers have discussed the idea of incorporating new and uncorrelated data and signals ever since 2007 quant crisis. However, traditional factors still overrule multifactor model world. The perception is that unconventional data tends to have short history, poor coverage, limited capacity, and is prone to data mining.
- **Sophisticated Modeling Techniques.** Ordinary least squares regression, linear multi-factor models, and mean-variance optimization were the main workhorses by quantitative managers prior to 2008. In recent years, nonlinear models and portfolio construction beyond mean-variance have received warm acceptance in the industry. Machine learning, despite of its wide adoptions in many other industries, is still in its infancy in institutional investing.
- **Smart Beta, Risk Premia, and Factor Investing.** One of the most significant (and controversial) developments in the investing world in recent years is undoubtedly factor investing. Asset management firms, banks, and index providers are all rushing to produce their own versions of smart beta indices. Rather than relying on portfolio managers to pick and mix factors, now investors can choose their own factor portfolios – in either ETF or structured products format – and conduct their own asset allocation.
- **Diversifying Globally.** As it becomes more and more difficult to add alpha in the US (and Japan), managers have diversified into other regions, e.g., Europe, UK, Canada, Australia, and emerging markets.

THE FUTURE

The battle between active and passive investing will only intensify in the future. We believe that the two survivors in the active investing business are likely to be:

- **Simple, Transparent, Cost Efficient, and Large Capacity Products.** The classic example in this category is obviously the traditional index products. A more recent example is the idea of smart beta. As factor investing has taken market shares away from pure market capitalization indices, liquid alternative products are gaining at the expense of hedge funds and private equities.
- **Sophisticated, Complex, and Limited Capacity Products.** The next frontier of leading edge investment is likely to be heavily involved in Big Data and machine learning – two topics that we have been emphasizing since 2006 and will continue to focus in the future.

HOW ARE WE DIFFERENT?

In a fairly mature industry, such as the active investing business, competition is fierce, alpha is scarce, and consistent outperformance is extremely difficult. More importantly, for sell-side research, how to add value to our clients – asset managers and asset owners – has always been a formidable task.

We have the luxury of re-building our entire infrastructure from scratch, using the latest technology and collective experience that we have accumulated over the years (but never really gotten the time to implement in the past).

Best-in-class Technology Infrastructure

We have developed and implemented a highly flexible and scalable cloud-based technology infrastructure. The entire system, from data integration, factor construction, single factor backtesting, multi-factor model development, portfolio simulation and implementation, transaction cost analysis, to performance attribution is developed in-house. As many of our models are computationally intensive, we are taking advantage of parallel computing, by leveraging hundreds of Linux servers.

Big Data Big Contents

We have formed a close relationship with both many traditional and unconventional data vendors. The content of our Big Data infrastructure ranges from traditional company fundamentals, market data, and global macroeconomics to unconventional/unstructured textual, images, people, and products. Introducing the most interesting and relevant vendors to our clients is one of our major mandates.

Highly Sophisticated Modeling Techniques

The team uses highly sophisticated modeling techniques from econometrics, time series analysis, computational statistics, machine learning and data mining, natural language processing (NLP), etc. The computing is done on a large suite of parallel Linux servers.

Combining Macro and Micro Research

The investment world has long been divided into macro investors and stock pickers (and similarly, fundamental and quantitative approaches). We aim to bring top-down macro and bottom-up stock selection together across our technology infrastructure and research.

Adding Alpha along the Entire Value Chain

There are many things that we can further refine along the entire quantitative investing process. While most managers and research analysts tend to focus on signal research, we conduct detailed and innovative research on every aspect. In the coming months, we expect to release a series of research on data, factor research, modeling techniques, risk and portfolio construction, and portfolio implementation.

THE QUANTITATIVE INVESTMENT PROCESS

The quantitative investment process is a remarkably complex system. Managers can add significant value in almost every step in the process. In an increasingly competitive field, even small improvements (id implemented properly) can still make a big difference in the end. In this and a sequence of upcoming papers, we will briefly explain the common approaches, best practices, and our own interpretations for each process. This paper covers the first two topics – infrastructure and data modeling.

- Big Data Infrastructure
  - From Hardware to Software
  - Big Data Contents
- Data Modeling
  - Data Error Checking
  - Outlier Control and Removal
  - Data Transformation
  - Missing Value Imputation
- Signal Research
  - Common Style Categories
  - In-house Developed versus Outsourced (e.g.,, alpha capture)
  - Factor Selection
- Conventional Multifactor Models
  - Equal Factor Weighting
  - Global Minimum Variance Weighting
  - Risk Parity Weighting
  - Grinold & Kahn Mean Variance Optimization
- Advanced Multifactor Models
  - Factor Weighting with Downside Risk and Tail Dependence
  - Factor Timing, Style Rotation, and Macro Overlay
  - Machine Learning Techniques
- Risk Models
  - Sample and Shrinkage Covariance Matrix
  - Fundamental Risk Models
  - Macroeconomic Risk Models
  - Statistical Risk Models

  - Hybrid Risk Models
- Portfolio Construction
  - Naïve Weighting (e.g., equally weighted, inverse volatility weighted, stratified sampling, etc.)
  - Mean-Variance Optimization (MVO)
  - Other Optimization Techniques (e.g., Risk Parity, Maximum Diversification, Minimum Tail Dependence, Mean-CVaR)
  - Portfolio Constraints: Turnover, ADV, Shorting, Market Neutrality, Country/Sector/Beta/Style Neutrality, etc.
- Transaction Cost Analysis (TCA)
  - Linear Transaction Cost Models
  - Quadratic Cost Models
  - Other Empirically Fitted TCA Models
  - Portfolio Optimization with Transaction Costs
  - Optimal Trade Execution
- Portfolio Implementation
  - Model Portfolios
  - Choice of Rebalancing Frequency
  - Choice of Trading Benchmark (e.g., Implementation Shortfalls, VWAP, etc.)
  - Choice of Trading Venues (e.g., Program Trading, Electronic Execution, etc.)
- Performance Attribution
  - Return Based
  - Holding Based
  - Custom Attribution
  - Active Hedging

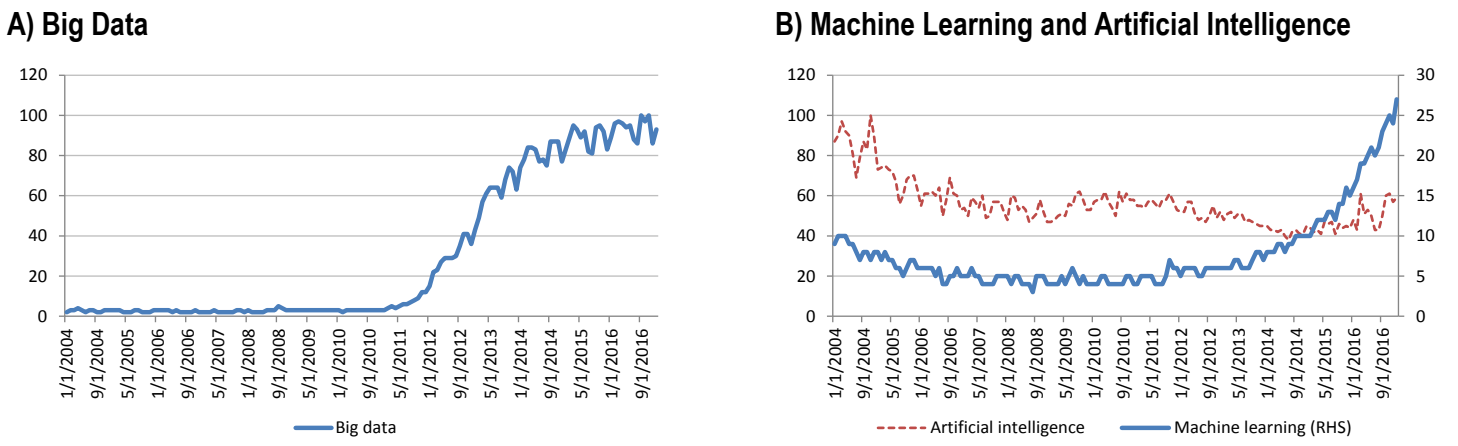
BIG DATA INFRASTRUCTURE

In the investment business, almost everybody would admit that data and technology infrastructure are critical to the long term success of the business. However, in practice, it is still generally being perceived as the least glamourous part of the firm. Portfolio managers and research analysts lack motivation to deal with data and technology, while the technology team does not necessarily understand the real business needs. We have seen so many cases of prolonged development cycle and even outright abandoning years of development efforts. In this section, we provide an overview and some concrete examples on this important topic.

The phrase "Big Data" has gained tremendous popularity in recent years (see Figure 13). We define Big Data as three key and related elements:

- **The underlying technology infrastructure** – both hardware and software. We need hardware to store and process our data. We need software to access, to analyze, and to extract patterns from our data.
- **The contents** – what we want to store and access. We have both structure and unstructured data. Data can be related to any entity. An entity can be a country, a government agency, a person, a security (a fixed income security, a tradable commodity, a currency, or more commonly, a public company). An entity may have multiple layers. For example, equity analysts primarily care about public companies, but a company may have multiple classes of stocks, trading on multiple exchanges, listed in multiple different countries, and denominated in multiple currencies.
- **The analytical toolbox** – what tools we can use to process the data contents. We can analyze our data either "manually" (i.e., the traditional Excel spreadsheet or canned statistical packages) or "automatically" (e.g., sophisticated machine learning algorithms).

Figure 13 Google Trends Interest in Big Data and Machine Learning



Sources: Google Trends, Wolfe Research Luo's QES

TECHNOLOGY INFRASTRUCTURE

For asset managers, it has been a constant debate whether they should build or rent the analytical system. Portfolio managers could build infrastructure, backtesting engine, risk and portfolio construction tools, portfolio management and performance attribution system, all or partially on their own. Alternatively, they can "rent" or subscribe one of the many (and growing) off-the-shelf solutions on the market, e.g., Factset, ClariFi and Market QA. There are significant trade-offs.

The two approaches are not mutually exclusive. We can certainly pick and choose the mission critical components to build on our own, while at the same time outsource the other elements to vendors. Figure 14 shows the pros and cons of these two approaches.

Figure 14 Comparing the Build or Rent Models

Consideration	Build your own	Rent from vendors
Development Cycle	Long	Short
Development Cost	High	Low
Development Risk	High	Low
Expertise	High	Low
Maintainance Cost	High	Modest
Operational Risk	Users need to ensure data is properly backed up, system is well maintained, and models are free from coding errors	The key operational risk is that one of our vendors go out of business
Learning Curve	In-house developed systems tend to be less user friendly with limited GUI (Graphical User Interface); therefore, the learning curve is steep	Most off-the-shelf products are more likely to be "point-and-click"; therefore, are easy to learn and use
Flexibility	It is far more flexible, as users can integrate new databases, test new modeling techniques, and integrate new systems at their own pace	Limited by vendors
Scalability	If a system is well designed, it should be scalable for larger data storage, more computing power, more users, and more locations.	Well respected vendors generally have scalable platforms
Speed and Power	A well developed system should allow parallell computing and efficient algorithms	Most off-the-shelf products are more likely to be "point-and-click"; therefore, not designed for batch processing
Sophistication	Users can add their own algorithms and models easily to their own system to ensure their research is cutting edge	Most vendor products offer some levels of customization, in that users can incorporate their own coding and models, but it is really limited by the vendors

Sources: Wolfe Research Luo's QES

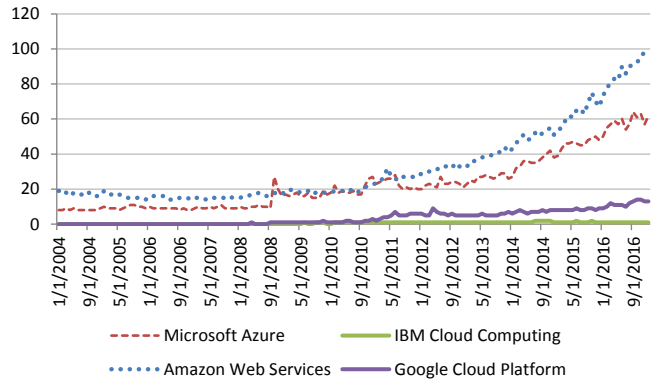
We decide to build the entire technology infrastructure on our own. However, unlike the traditional data warehouses that reside on a bank's own data centers, we are able to take advantage of the far more powerful and flexible cloud structure. There are huge advantages of storing our databases on the cloud. Today's public clouds (e.g., Amazon, Google, Microsoft, and IBM) and growing numbers of private clouds are taking more and more market shares from traditional corporate data centers, for a number of good reasons (see Figure 15):

- It is far more cost efficient. The pay-as-you-go model means we only pay for the storage and computing power of our needs, rather than having substantial excess back-up capacities.
- It is much faster to deploy and expand. Initiating our data warehouse takes a matter of days rather than months, because we do not need to purchase multiple servers.

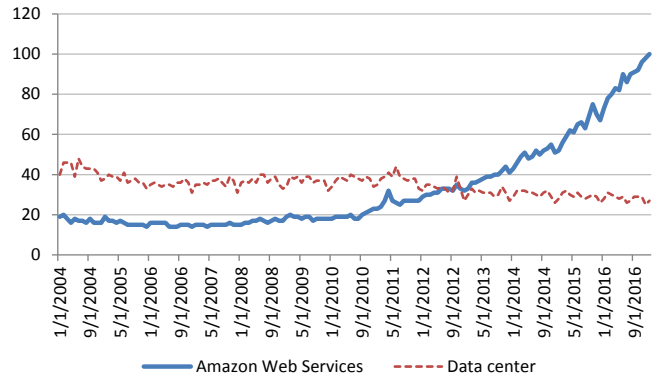
- It is significantly more scalable, as adding additional storable and computing power is normally just a few clicks away. This is crucial, as the size of our current database is already hundreds of terabytes and it is growing rapidly.
- Back-up and data security. It is still somewhat controversial of whether cloud databases are fully backed up and as secure as corporate data centers. In our experience, it appears that the largest cloud providers have fairly robust data back-up and security measures in place.
- Big Data support. Cloud infrastructure supports Big Data technologies, such as scalable distributed file system (S3, Hadoop) and MapReduce framework.
- Parallel computing. Many of our models are computationally intensive. Running a single period random forest model can easily take a few hours. To conduct a complete backtesting over 30 years means that we have to wait for a month, using our desktop computer. With parallel computing, we can divide the job into hundreds of servers and finish the backtesting overnight.

Figure 15 Google Trends Interest in Cloud

A) Cloud



B) Data Center



Sources: Google Trends, Wolfe Research Luo's QES

Back-End Data Storage

On our cloud, we use a hybrid of Oracle and Microsoft SQL Server databases and data warehouses hosted on a suite of Linux Ubuntu servers.

We have access to a large number of data vendors. However, rather than using a third-party data aggregator, we access vendors' data directly, mostly via FTP and flat files. We then load the raw data into our databases.

The key philosophy of relational databases is that data is highly normalized to save storage space and ensure data integrity. However, the downside of a normalized database is that it is rather counterintuitive for data scientist. Researchers are used to deal with matrix-type of two-dimensional data[10]. To retrieve data in a user-friend format, analysts would have to join multiple tables via

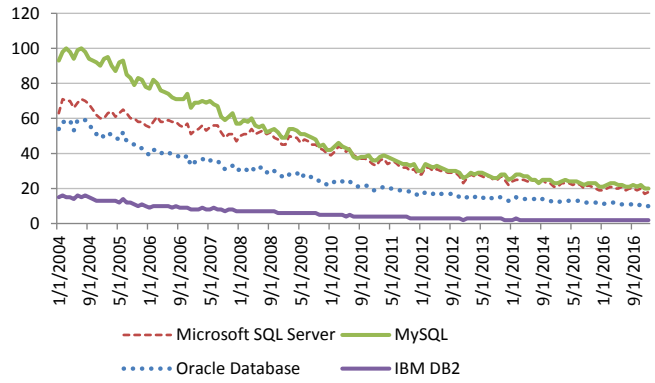
[10] Some problems are more naturally fit into multi-dimensional arrays. For example, when we deal with institutional ownership data, we need to model companies (i.e., stocks), time, and institutional owners (i.e., funds) at the same time. In data modeling, however, multi-dimensional arrays are normally transformed to matrices first, as most modeling techniques are designed to handle matrix data.

complex SQL queries. As relational database is gradually losing market share (see Figure 16 A) and to efficiently use SQL requires extensive experience[11], it often seems to be a difficult task for analysts to use relational databases directly. It is also terribly prone to errors. In our experience, this is the primary reason why most portfolio managers use third-party data aggregators (e.g., Factset or ClariFi) rather than building their own databases.

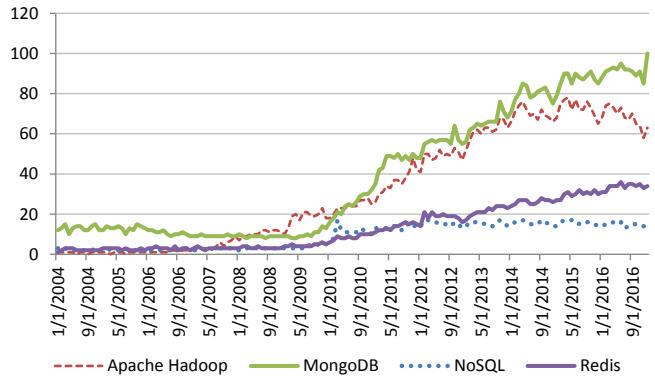
The rapid increase in the size and complexity of Big Data has also spurred the fast development in alternative database structures. One example is the introduction of distributed file system (e.g., Hadoop), where databases are essentially a collection of files. Files can be distributed over multiple systems. A processing engine can query all these files and provide analytics to a client. This architecture is set up for immensely large datasets that require speedy access. Other alternatives include document oriented noSQL databases such as MongoDB. Figure 16 (B) shows the level of interest on these unconventional databases has surged in recent years.

Figure 16 Google Trends Interest in Database

A) Traditional Relational Databases



B) Trends in noSQL Databases



Sources: Google Trends, Wolfe Research Luo's QES

Middle Layer Data De-normalization

The innovative structure that we have developed is to construct a middle layer, we internally codenamed LQuant. The LQuant is written in a combination of SQL, Java, Python, and R. It automatically de-normalizes raw data into two-dimensional matrices (or multi-dimensional arrays) that analysts can use directly, without any knowledge of SQL.

To build the LQuant structure, however, takes extraordinary efforts. Developers need to have a thorough understanding the exact schema of every single underlying database and more importantly, need to know precisely how research analysts would like the data to be structured.

Front End Data Analytics

In practice, most analysts and portfolio managers are only exposed to the front-end system. It can be a user-friendly GUI (Graphical User Interface) such as Factset and ClariFI with point-and-click

[11] We find most computer science programs at universities only expose students to modest relational database and SQL. Most computer science graduates seem to lack hands-on SQL experience these days.

access. Alternatively, it can be written in statistical languages such as R and Python. Our front end system is also rather complex with multiple modules:

- A general purpose backtesting engine (LBacktester). As we will discuss in Part II (*Signal Research and Multifactor Models*) of this research series, the first step of modeling is often to replicate a realistic investment process, back in time, and track the performance.
- A suite of multi-factor modeling techniques (LMultiFactor). Once we have the raw ingredients, i.e., factors, we need to build multifactor models. There are many ways of choosing signals and weighting them properly. Advanced multifactor models involve factor timing and machine learning techniques, which will be addressed in Part III (*Style Rotation, Machine Learning, and the Next Frontier of Systematic Investing*).
- A risk, transaction cost, and portfolio construction platform. Transforming from alpha signals to a real-life portfolio requires us to balance among expected return, risk, correlation (i.e., diversification benefit), transaction costs, liquidity and short availability, and other institutional constraints.
- A portfolio attribution system. Lastly, we want to understand the sources of risk and returns in our portfolio. Both portfolio construction and attribution will be covered in Part IV (*From Theory to Practice*).

The Choice of Programming Languages

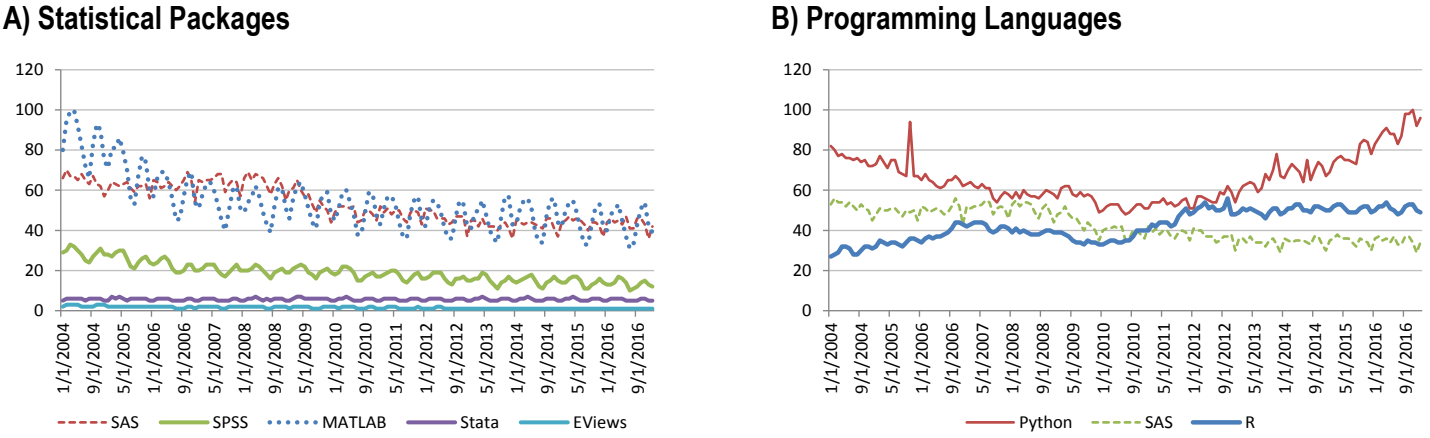
Another common question that we hear from clients all the time is what programming language we should use. For the purpose of this discussion, we focus on the coding language for the front-end backtesting and research systems.

Traditional statistical computing languages such as Matlab and SAS have been continuously losing market share (see Figure 17 A). On the other hand, free, dynamic and cutting edge open source languages such as R and Python have been gaining prominence (see Figure 17 B).

Traditional languages offer strong support, stable environment and development tools, while open source languages offer more cutting edge algorithms. Latest functionalities and methods get added quickly in the open source world, while it takes years to implement by the traditional canned packages. There are also challenges with open sources languages, such as:

- There are often many libraries and functions for the same task, while choosing the “best” one is not always straightforward;
- Documentation of functions and examples is generally weak;
- Technical support lags behind commercial vendors; and
- Not all libraries and packages are updated and synchronized when the main language is upgraded.

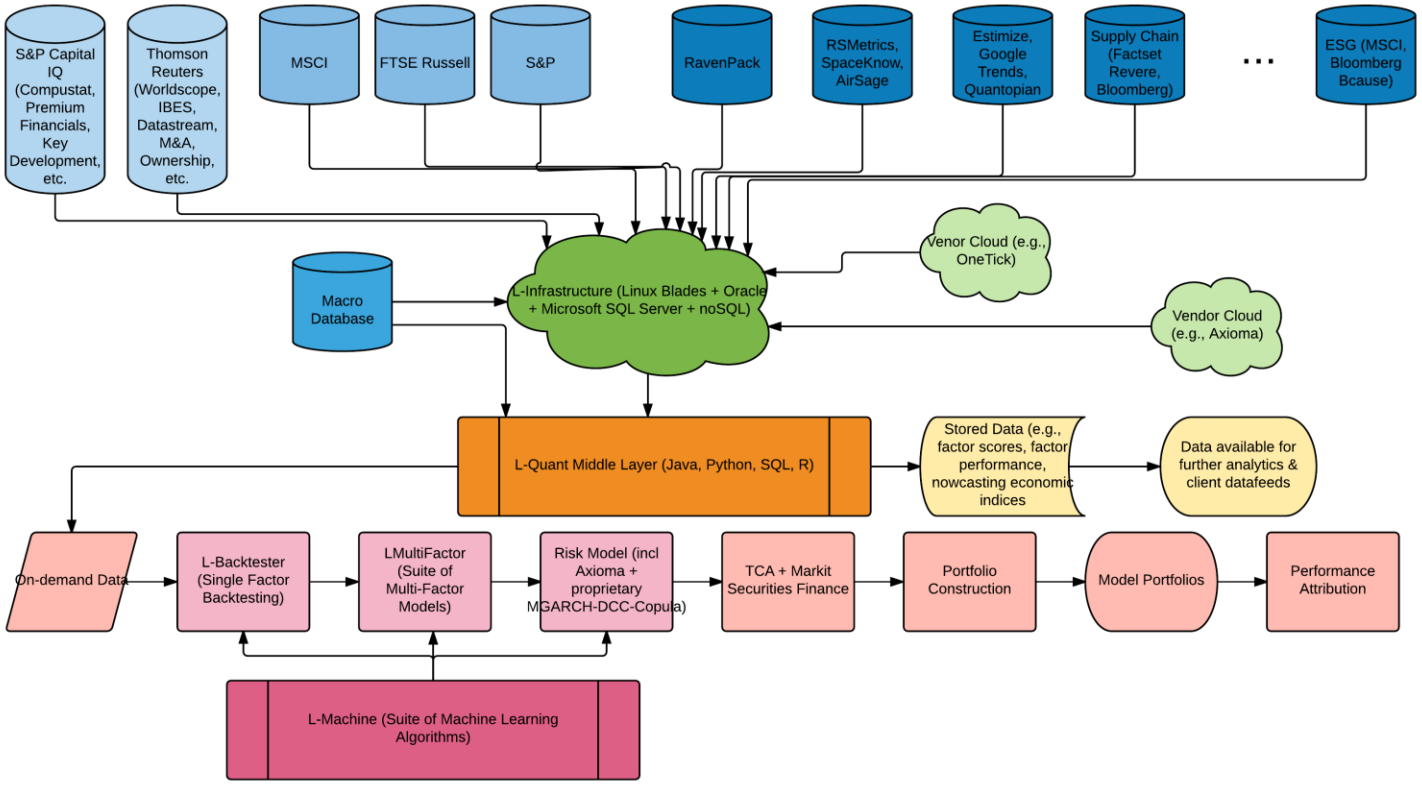
Figure 17 Google Trends Interest in Programming Languages and Statistical Packages



Sources: Google Trends, Wolfe Research Luo's QES

Figure 18 shows our overall technology infrastructure.

Figure 18 The Three Components of Big Data



Sources: Wolfe Research Luo's QES

RESEARCH UNIVERSE

Our global equity research universe covers the vast majority of listed stocks (both large- and small-cap) in all developed and emerging market countries.

Geographically, we use all countries[12] in the MSCI All Country World Index (MSCI ACWI). For the US market, we use all stocks in the Russell 3000 index. For Canada, we use the S&P/TSX Composite Index and for all other countries, we use the S&P BMI (Broad Market Index) as our research and investment universe.

For most of our research, we divide the world into ten regions:

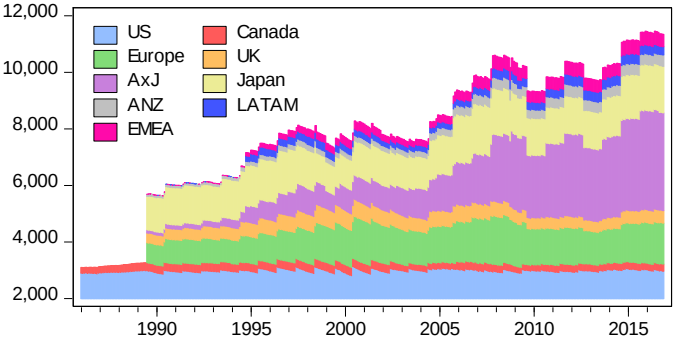
- US (Russell 3000 universe)
- Canada (S&P/TSX Composite universe)
- Europe ex UK (S&P BMI universe)
- UK (S&P BMI universe)
- Asia ex Japan (S&P BMI universe)
- Japan (S&P BMI universe)
- Australia and New Zealand (S&P BMI universe)
- LATAM (S&P BMI universe)
- Emerging EMEA (S&P BMI universe)
- China A (MSCI China A universe) – to be launched in the future

As shown in Figure 19 (A), US, Europe, AxJ, and Japan are the four largest regions by number of stocks, while US market counts for half of the global equity market by market capitalization (see Figure 19 B). Sector-wise, industrials, consumer discretionary, financials, and information technology have the most numbers of stocks and largest market capitalization (see Figure 19 C and D).

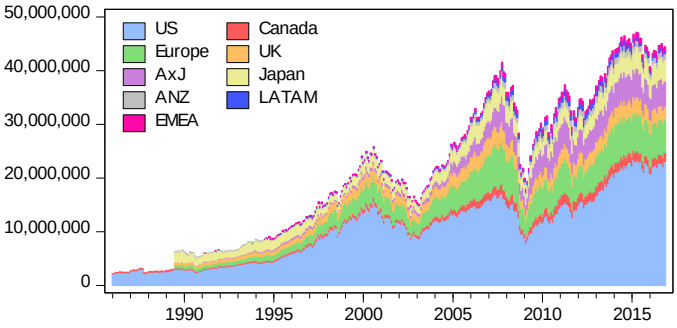
[12] The number of countries in the MSCI ACWI changes from time to time. Currently, there are 23 developed countries and 23 emerging markets in the index, for a total of 46 countries.

Figure 19 The Global Investment Universe, by Geographic Regions

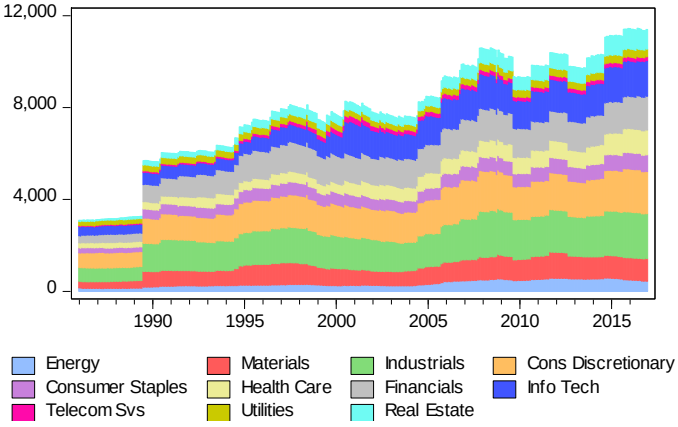
A) # of Stocks, by Region



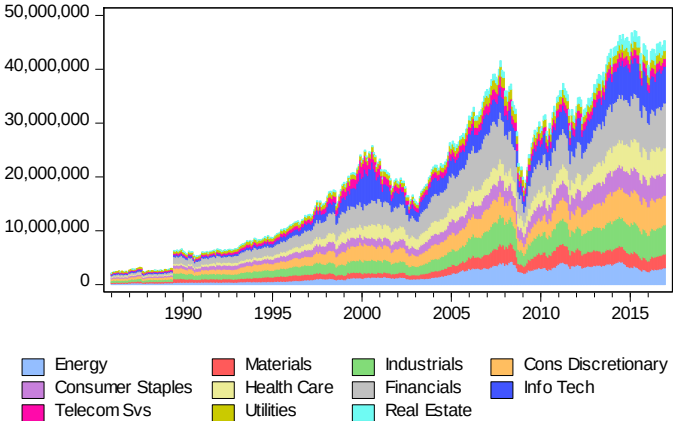
B) % of Market Capitalization, by Region



C) # of Stocks, by Sector



D) % of Market Capitalization, by Sector



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

Correlation and Opportunities

The equity markets in most regions are somewhat correlated (see Figure 20 A), with the exception of Japan[13]. For stock pickers, the cross-sectional return dispersion is another useful metric. US and AxJ not only offer the greatest breadth (i.e., number of stocks), but also the largest return dispersion (see Figure 20 B).

Among the 11 GICS sectors, energy and utilities are slightly less correlated to other sectors (see Figure 20 C). Health care and information technology stocks show the largest dispersion, while stock picking offers relatively fewer opportunities in the utilities space (see Figure 20 D).

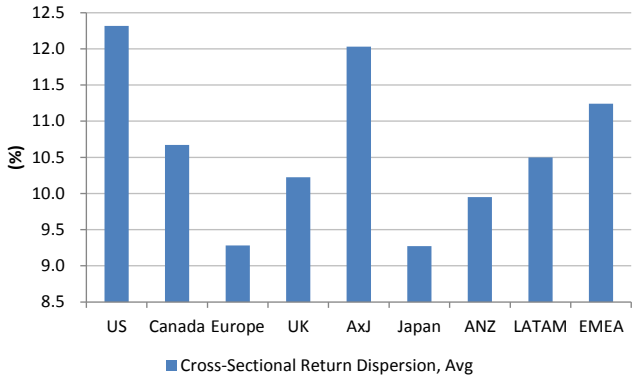
[13] The correlation matrix is based on the equally weighted index for each region. Capitalization weighted indices generally have higher correlation. Since we are more interested in stock selection, equally weighted indices are more relevant.

Figure 20 Correlation and Opportunities around the World

A) Correlation across Regions

	US	Canada	Europe	UK	AxJ	Japan	ANZ	LATAM	EMEA
US	100%								
Canada	77%	100%							
Europe	77%	72%	100%						
UK	77%	71%	88%	100%					
AxJ	62%	65%	61%	60%	100%				
Japan	42%	36%	45%	45%	41%	100%			
ANZ	68%	71%	71%	71%	63%	43%	100%		
LATAM	63%	69%	70%	67%	68%	37%	65%	100%	
EMEA	68%	71%	77%	76%	66%	43%	69%	75%	100%

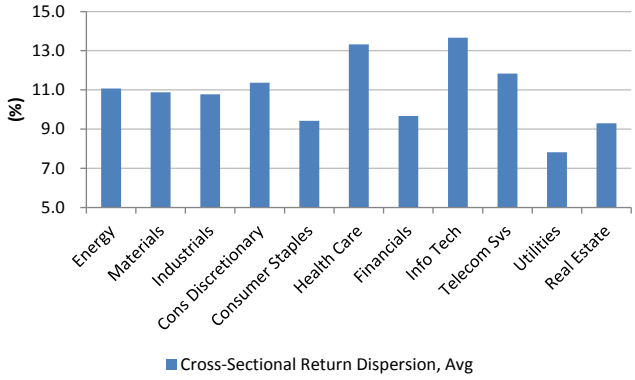
B) Cross-sectional Return Dispersion, by Region



C) Correlation among Sectors

	ENRS	MATR	INDU	COND	CONS	HLTH	FINL	INFT	TELS	UTIL	RLST
Energy	100%										
Materials	76%	100%									
Industrials	71%	93%	100%								
Cons Discretionary	65%	88%	96%	100%							
Consumer Staples	64%	86%	93%	92%	100%						
Health Care	54%	69%	81%	83%	79%	100%					
Financials	65%	83%	91%	92%	89%	79%	100%				
Info Tech	52%	71%	81%	83%	69%	82%	74%	100%			
Telecom Svs	47%	62%	69%	75%	63%	72%	70%	86%	100%		
Utilities	63%	70%	74%	70%	78%	65%	78%	53%	55%	100%	
Real Estate	63%	84%	86%	85%	84%	69%	86%	64%	57%	75%	100%

D) Cross-sectional Return Dispersion, by Sector



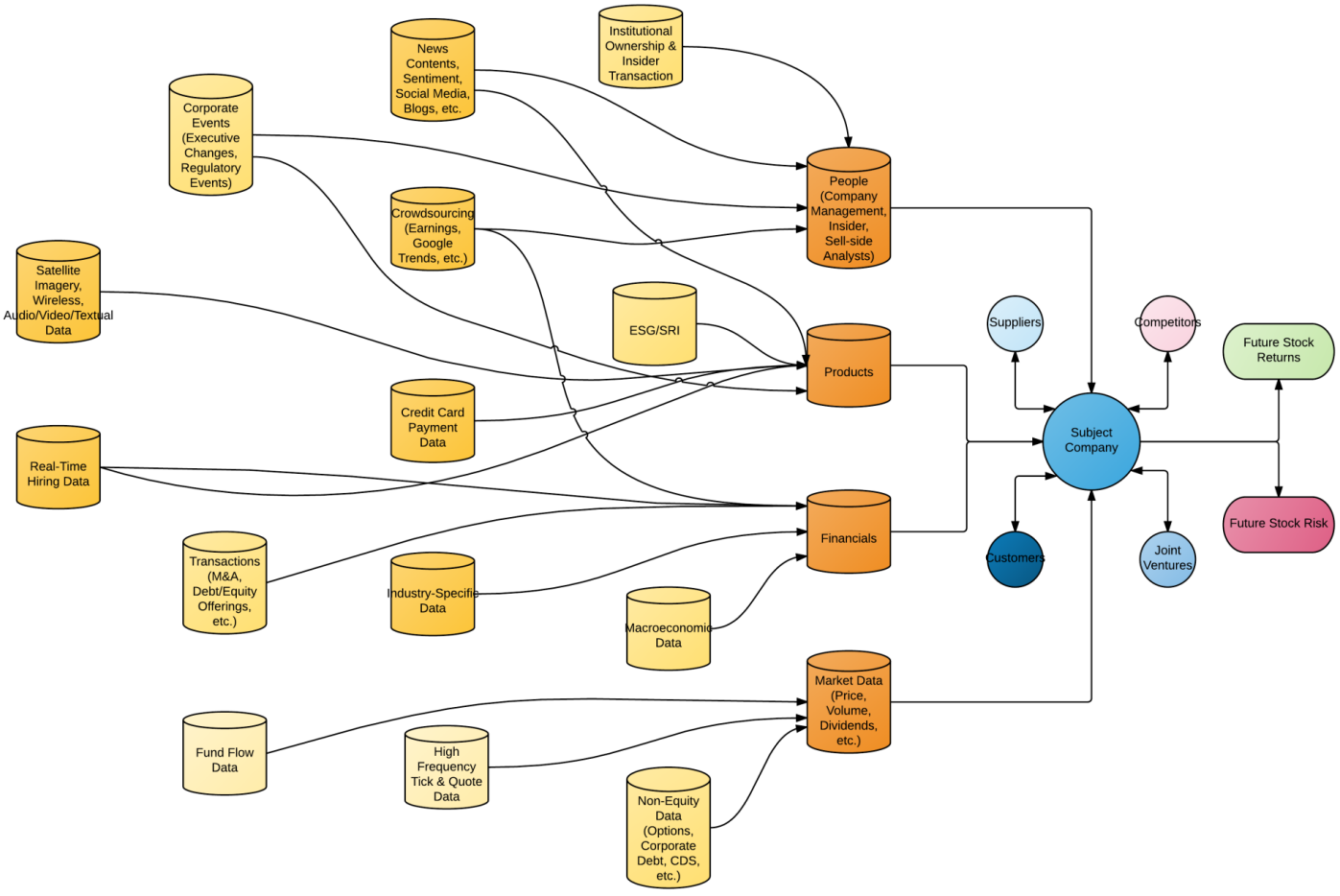
Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

CONTENTS – THE BIG DATA EVOLUTION

Our team has access to a wide range of both traditional and highly unique data sources, including global company fundamental data, pricing and market data, major global indices, global economic data, corporate events, business relationship, institutional ownership and insider trading, product and professional, M&A and other transactions, satellite image and mobile phone locations, textual and news sentiment, etc.

With our subject company in mind, we want to predict the return and risk of the target stock. A company is linked to its customers, suppliers, competitors, and other related entities. We can analyze a company along four dimensions – people (e.g., executive management, board members and other insiders, sell-side analysts covering the company, institutional shareholders and creditors), products and services (that the company makes or provides, compliment and substitute products), financials (GAAP, IFRS, and regulatory filings), and market data (including corporate events). The vast array of data and related entities form an immensely complex and exciting web of data and opportunities (see Figure 21).

Figure 21 The Complex Web of Big Data



Sources: Wolfe Research Luo's QES

Entity Relationship and Master Mapping Tables

For global stock selection modeling, we need to have three basic layers to uniquely identify a specific issue of a stock: company, issue, and exchange. The company level identifies a specific public company, e.g., IBM or Samsung. A public company may have multiple issues of equity securities. For example, Google[14] has three classes of equity securities: Class A (exchange ticker GOOGL) shares with one vote per share, Class B (not publicly traded, 10 votes per share), and Class C (exchange ticker GOOG) with no voting rights. Both GOOGL and GOOG are in the S&P 500 and Russell 1000 indices. Finally, even the same issue/class of equity securities can be listed and traded on multiple stock exchanges – the exchanges can be in different countries and the stocks can be traded in different currencies. For example, Deutsche Bank has only one class of shares. However, the bank's stocks are traded in both Germany (on the Frankfurt Stock Exchange, under the ticker of DBK) and in the US (on the New York Stock Exchange, with the ticker of DB).

For our modeling purpose, we use a combination of company + issue + exchange to uniquely identify a specific stock. The most important message is that almost all common identifiers, e.g., company name, exchange ticker symbol[15], and even CUSIP and SEDOL can change, due to corporate actions. Therefore, to avoid look-ahead and survivorship biases, we need to have a proper point-in-time identifier for each company, each issue, on each exchange, at each given point in time. Furthermore, for each and every database we use, for all historical data, we need to have the same consistent point-in-time identifiers or find other ways to map them consistently (see Figure 22 for an example). It is a rather daunting task. However, as we will show in the "Survivorship Bias" section, without the ability to track companies point-in-time can result serious problems in our backtesting and lead to a disastrous investment strategy.

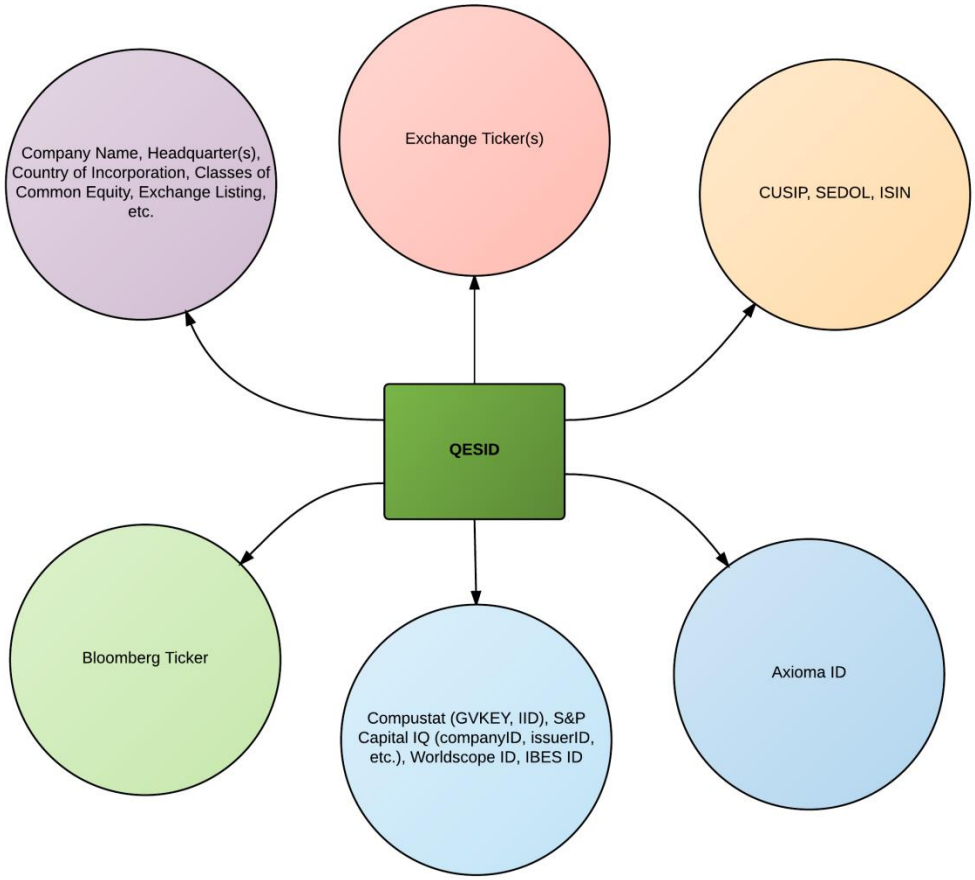
As we move from stock selection to global macro, the entity mapping issue becomes even more complex. For example, a company's stock can be incorporated in one country, while its stocks can be traded in another country (or countries). It may operate in many countries, while its suppliers and customers can be in the same or different sets of geographic regions. Companies have alternatives when they conduct currency translations on their financial statements and they can engage in very different kinds of currency hedging (e.g., fair value hedge versus cash flow hedge).

There are a few dominant vendors providing entity mapping:

- Thomson QA
- S&P Capital IQ
- Factset
- Axioma

[14] Please note that the legal name of Google is actually Alphabet Inc.
[15] Ticker change and company legal name change are actually interesting stock selection factors.

Figure 22 Mapping Companies and Issues across Vendors and over time



Sources: Wolfe Research Luo's QES

Market Data

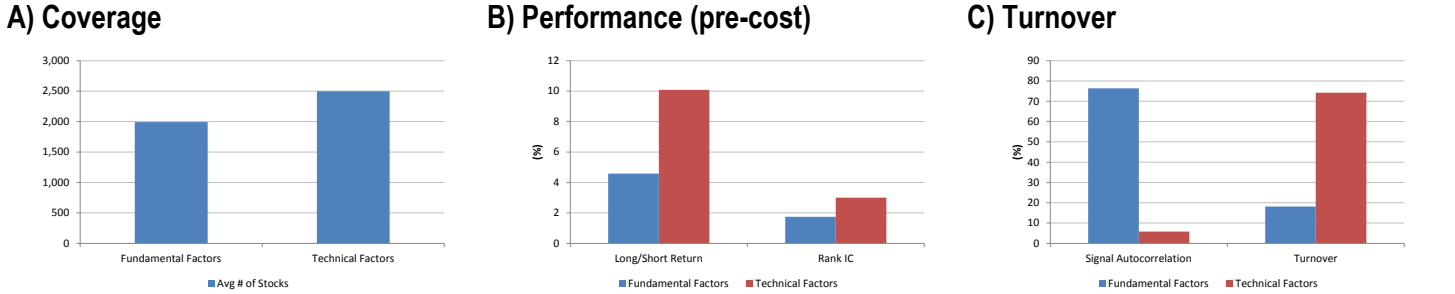
Market data refers to stock price, dividend, split, number of shares outstanding, and other corporate action data. Market data is critical for stock return, liquidity, and market capitalization calculation. Most technical factors also rely on accurate market data to compute. Traditional market data is available on a daily (or lower) frequency. High frequency trading firms, market making business, and transaction cost analytics typically need to access tick-and-quote (TAQ) database. TAQ databases have a simple structure (e.g., time stamp, transaction quantity, prevailing bid and ask prices, transaction price, and limit order book), but the size of database is gigantic, which poses severe challenges to data storage and processing. In our previous research, we use specialized database (i.e., KDB) and programming language (KDB+ and Java) to process high frequency tick-by-tick data (see Webster, et al [2015]).

We have a large number of technical indicators in our factor library. On average technical factors have much better coverage[16] (see Figure 23 A), stronger pre-cost performance (see Figure 23 B), but suffer from higher extraordinarily turnover (see Figure 23 C) than signals based on fundamental data.

[16] Technical factors are primarily based on market data, e.g., price and volume data, which is available to almost all companies.

In Part IV of this research series, we will show the implication of transaction costs on model performance and discuss how to incorporate costs in our portfolio construction.

Figure 23 Technical versus Fundamental Factors



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

Common market data vendors include:

- S&P Capital IQ, Compustat, IDC
- Thomson Reuters and Datastream
- CRSP
- Markit Securities Finance
- Tick Databases: One Tick, Thomson Reuters, Tick Data

Company Fundamental Data

Public companies in almost every country are required to disclose their financial performance via periodic financial statements and other filings with local and sometimes also foreign regulators. There are three main sets of financial statements: balance sheet, income statement, and statement of cash flows. Companies report their financials on different frequencies – quarterly, tri-annually, semi-annually, and annually. Currently, there are two major accounting standards – US GAAP and IFRS[17]. Companies have great deal of flexibility on how they want to present their financial performance, while data vendors attempt to reconcile the differences into one consistent template globally across industries.

In addition to the three main sets of financial statements, there are also other important disclosures contained in footnotes, MD&A (Management Discussion & Analysis), and other filings.

There are a few vendors providing fundamental data:

- S&P Capital IQ: Compustat, Capital IQ Premium Financials
- Thomson Reuters: Worldscope, Reuters Fundamentals
- Bloomberg
- Factset

[17] A few major countries have their own local accounting standards (e.g., China and Japan), but they are largely in line with IFRS.

When we choose data vendors, we need to take into account of the following: history, coverage, updating timeline, point-in-time structure, database design, data delivery mechanism, and most important of all, data quality.

For North American companies in the US and Canada, the benchmark data is Compustat North American package. Compustat has a point-in-time database called Compustat Snapshot North America and that is where we source our data.

For companies outside of the US and Canada, we use a hybrid structure. We combine Capital IQ's Premium Financials (CIQPF) and Worldscope databases. CIQPF is a point-in-time database with extensive coverage of most financial statement and supplemental data items, in both annual and interim frequencies (depending on a company's own reporting frequency). Worldscope has the longest history and covers the most number of companies, especially back in history. However, Worldscope data is not point-in-time[18] and its interim data is more limited.

Although the interest of academic and practitioner's research has shifted away from financial statement based factors in recent years, there is still meaningful untapped alpha in this space. For example, Figure 24 shows the income statement template from three main data vendors we use (Compustat, S&P Capital IQ Premium Financial, and Worldscope). It is probably surprising to many investors to see how different it can be, even for the most well-known income statement. We hardly ever see any discussion of or attempt to reconcile the differences across vendors in the existing research literature.

[18] Thomson Reuters has a separate Point-in-time Worldscope database.

Figure 24 The Difference in Financial Statement Presentation – Income Statement Template

A) Compustat

Compustat Interim Income Statement Template

- Sales/Turnover (Net)
  - Sales/Turnover is net of Excise Taxes
- Operating Expense
  - Cost of Goods Sold
  - Gross Profit
  - Selling, General and Administrative Expenses
    - Operating Expense includes:
    - Research and Development Expense
    - Advertising Expense
- Operating Income Before Depreciation / EBITDA
- Depreciation and Amortization - Total
- Operating Income After Depreciation / EBIT
- Interest Expense
- Nonoperating Income (Expense) - Total
- Special Items
- Pretax Income
- Income Taxes - Total
  - Income Taxes include Income Taxes - Deferred
- Income Before Extraordinary Items and Noncontrolling Interests
- Noncontrolling Interest - Income Account
- Income Before Extraordinary Items
- Dividends - Preferred/Preference
- Income Before Extraordinary Items - Available for Common
- Common Stock Equivalents - Dollar Savings
- Income Before Extraordinary Items - Adj for Common Stock Equivalents
- Extraordinary Items and Discontinued Operations
  - Extraordinary Items
  - Discontinued Operations
    - Extraordinary Items include:
    - Accounting Changes - Cumulative Effect
- Net Income (Loss)

B) S&P Capital IQ

S&P Capital IQ Income Statement Template

- Revenue
- Other Revenue, Total
  - Gain(Loss) on Sale Of Assets (Rev)
  - Gain(Loss) on Sale Of Invest. (Rev)
  - Interest And Invest. Income (Rev)
  - Other Revenue
- Total Revenue
- Cost Of Revenue
  - Cost Of Goods Sold
  - Interest Expense - Finance Division
- Gross Profit
- SG&A Exp., Total
  - Selling General & Admin Exp.
  - Provision for Bad Debts
  - Stock-Based Compensation
  - Pre-Opening Costs
- R & D Exp.
- Depreciation & Amort., Total
  - Depreciation & Amort.
  - Amort. of Goodwill and Intangibles
- Other Operating Expense/(Income)
- Other Operating Exp., Total
- Total Operating Expenses
- Operating Income
- Interest Expense
- Interest and Invest. Income
- Net Interest Exp.
- Other Non-Operating Exp., Total
  - Income / (Loss) from Affiliates
  - Currency Exchange Gains (Loss)
  - Other Non-Operating Inc. (Exp.)
- EBT Excl Unusual Items
- Merger & Restruct. Charges
  - Restructuring Charges
  - Merger & Related Restruct. Charges
- Impairment of Goodwill
- Gain (Loss) On Sale Of Invest.
- Gain (Loss) On Sale Of Assets
- Other Unusual Items, Total
  - Asset Writedown
  - In Process R & D Exp.
  - Insurance Settlements
  - Legal Settlements
  - Other Unusual Items
- Total Unusual Items
- EBT Incl. Unusual Items
- Income Tax Expense
- Earnings from Cont. Ops.
- Earnings of Discontinued Ops.
- Extraord. Item & Account. Change
- Net Income to Company
- Minority Int. in Earnings
- Net Income
- Pref. Dividends and Other Adj.
- NI to Common Incl. Extra Items
- NI to Common Excl. Extra Items

C) Worldscope

Worldscope Income Statement Template

- New Sales or Revenues
- Cost of Goods Sold
- Depreciation, Depletion & Amortization
  - Depreciation
  - Amortization of Intangibles
  - Amortization of Deferred Charges
- Gross Income
- Selling, General & Administrative Expenses
- Other Operating Expenses
- Operating Expenses - Total
- Operating Income
- Extraordinary Credit - Pre-Tax
- Extraordinary Charge - Pre-Tax
- Non-Operating Interest Income
- Reserves - Increase/Decrease
- Pre-Tax Equity in Earnings
- Other Income/Expense - Net
- Interest Expense on Debt
- Interest Capitalized
- Pre-tax Income
- Income Taxes
  - Current Domestic Income Tax
  - Current Foreign Income Tax
  - Deferred Domestic Income Tax
  - Deferred Foreign Income Tax
  - Income Tax Credits
- Minority Interest
- Equity in Earnings
- After Tax Other Income/Expense
- Discontinued Operations
- Net Income Before Extraordinary Items/Preferred Dividends
- Preferred Dividend Requirements
- Net Income after Preferred Dividends
- Extraordinary Items & Gain/Loss Sale of Assets

Sources: Bloomberg Finance LLP, FTSE Russell, Markit, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

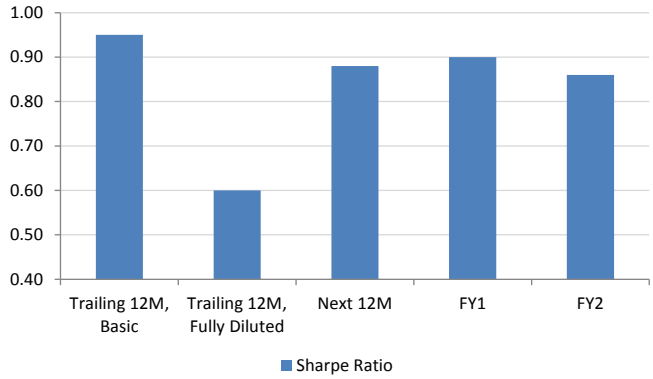
Let's take a simple example. We see reference to earnings yield (or price-to-earnings multiple) in academic and industry research all the time. However, few researchers elaborate on the exact metric of EPS used in the calculation. In practice, there are multiple variations of EPS that we can choose (see Figure 25 A) and the performance can be drastically different (see Figure 25 B). For example, the trailing earnings yield factor based on basic EPS has a Sharpe ratio almost 60% higher than the one based on fully diluted EPS.

Figure 25 Earnings Yield Factor, Based on Different EPS Metrics

A) Different EPS Metrics

	Basic	Diluted	Operating/ Normalized
Last Fiscal Year	√	√	√
Trailing 12M	√	√	√
Next 12M	√	√	√
FY1	√	√	√
FY2	√	√	√

B) Earnings Yield Factor Sharpe Ratio, US



Sources: Bloomberg Finance LLP, FTSE Russell, Markit, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

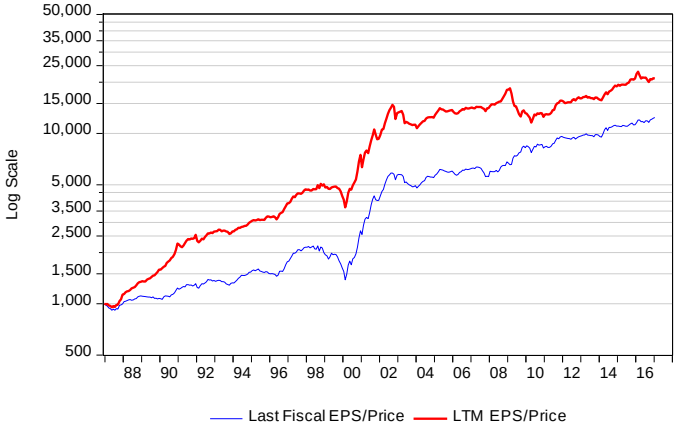
Another common issue on financial statement data is whether we should use interim data. Companies in the US and Canada almost all reported on a quarterly basis. Outside of the North American market, however, it starts to get complicated – companies can release their financials on a quarterly, semi-annual, tri-annual, or only annual basis. There are also firms that report on a nine-month or other irregular cycles. Furthermore, the reporting frequency of the same company may change over time. Similarly, a company may also change its fiscal year end. To show the impact of interim financial statement data, we backtest the performance of the earnings yield factor, with two variations:

- Last fiscal year annual EPS/Price
- Trailing 12M EPS/Price

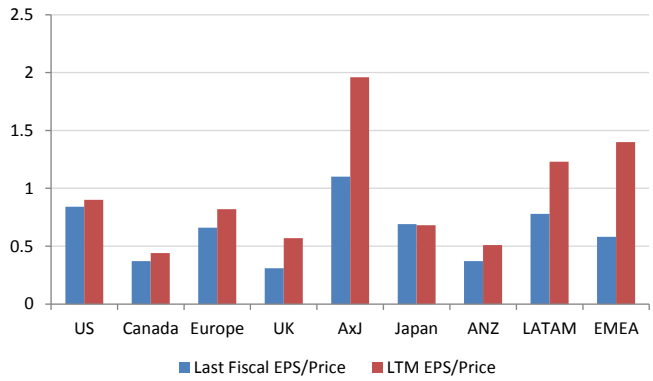
As shown in Figure 26 (A), the huge effort of incorporating interim financial statement data pays off handsomely. The cumulative performance of the earnings yield factor using interim EPS beats the same factor with last fiscal year EPS by almost 70% in 30 years. The boost in performance from more timely data is even more pronounced in AxJ and emerging markets such as LATAM and EMEA (see Figure 26 B).

Figure 26 The Incremental Benefit of Using Interim Financial Statements

A) Cumulative Performance, US



B) Sharpe Ratio, Global



Sources: Bloomberg Finance LLP, FTSE Russell, Markit, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

Industry-Specific Matters

The interest in Big Data seems to be centered on unconventional data sources. However, there is still tremendous untapped data in our traditional financial statements. We have heard the comment such as "financial statement data has been thoroughly studied and data mined in both academia and by practitioners; therefore, it is unlikely to find anything new". We would have to disagree. After an extensive academic literature review, for example, we find extremely limited publications on anything related to specialized industries such as banks, insurance companies, other financial services firms, and utilities. Ironically, most traditional data vendors such as Compustat, S&P Capital IQ and Worldscope actually have specific financial statement templates for these companies. It is just that few researchers have ever spent much time on these databases. Many analysts probably are not even aware of the existence of industry-specific financial statement formats.

Industry-specific accounting is not taught in most accounting programs at universities. Professional accountants tend to be organized by industries – only few specialize in industries such as banks and insurance. Investment analysts and managers rarely have a deep understanding of the idiosyncrasies of specialized industry accounting either.

Figure 27 shows a simplified bank balance sheet. For non-specialists, it clearly shows that bank financial reporting is quite different from industrial and consumer companies. The distinction of current and non-current assets is irrelevant. Investment securities and loans account for the vast majority of a bank's assets. On the liabilities and shareholders' equity side, banks tend to be substantially more leveraged than regular firms. The amount of equity capital is mostly mandated by regulators (which could be in multiple jurisdictions).

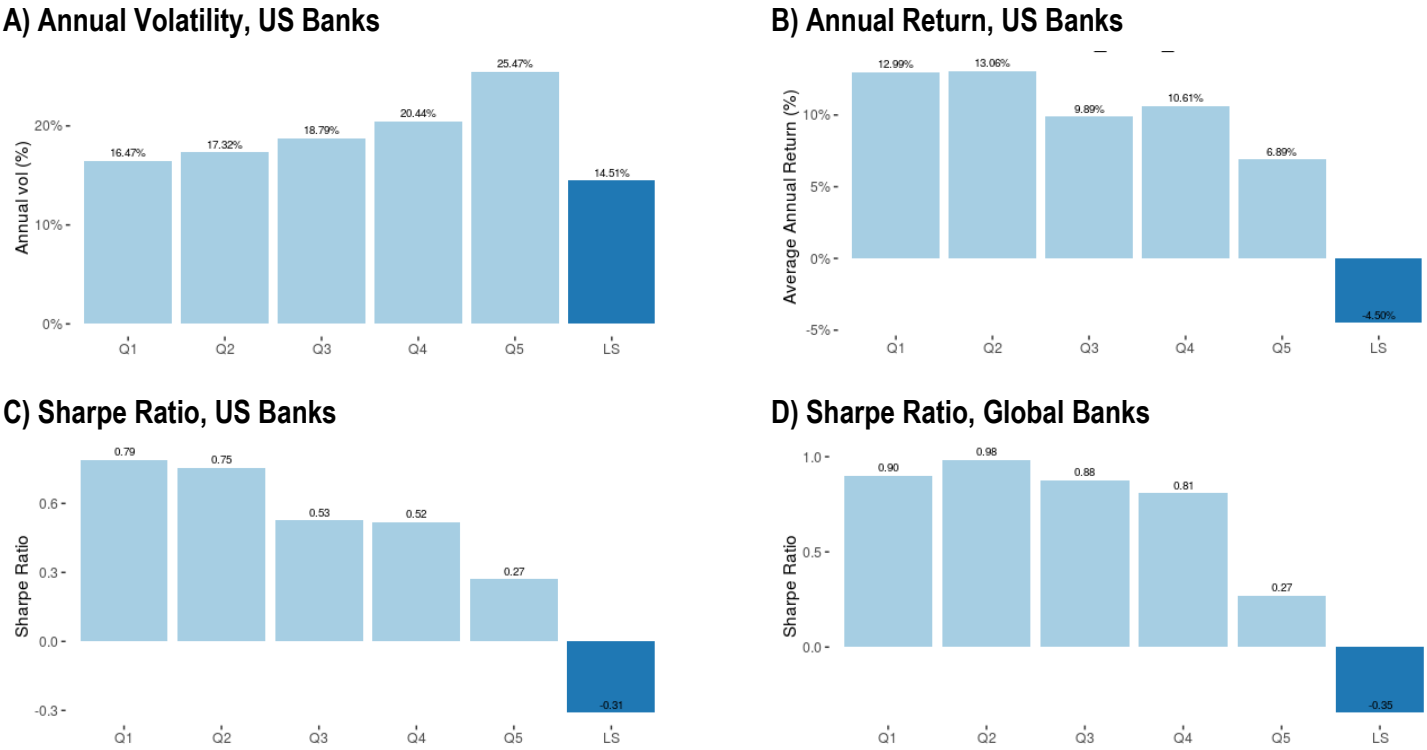
Figure 27 A Simplified Bank Balance Sheet

BALANCE SHEET	
Assets	*Liabilities*
Cash And Equivalents	Accounts Payable
Investment Securities	Accrued Expenses
Trading Asset Securities	Total Deposits
Mortgage Backed Securities	Short-term Borrowings
Total investments	Curr. Port. of LT Debt/Cap. Leases
Gross Loans	Long-Term Debt
Allowance For Loan Losses	Federal Home Loan Bank Debt - LT
Other Adj. to Gross Loans	Capital Leases
Net Loans	Trust Pref. Securities
Gross Property, Plant & Equipment	Other Liabilities, Total
Accumulated Depreciation	*Total Liabilities*
Net Property, Plant & Equipment	
Goodwill	*Shareholders' Equity*
Other Intangibles	Total Pref. Equity
Investment in FHLB	Total Common Equity
Other Assets, Total	Minority Interest
	Total Equity
Total Assets	*Total Liabilities And Equity*

Sources: S&P Capital IQ, Wolfe Research Luo's QES

We start from an example of bank-specific factor – loan loss provision ratio, defined as loan loss provision divided by operating income. Loan loss provision is an income statement item unique to banks. It is an estimated expense item set aside as an allowance for uncollected loans. Banks with high loan loss provision ratios are perceived as taking too much credit risk; therefore, their share prices may suffer. Based on our backtesting, banks with the highest loan loss provision ratios exhibit the highest equity volatility (see Figure 28 A), lowest return (see Figure 28 B), and therefore, the lowest Sharp ratio (Figure 28 C) in the US. We also observe similar patterns for global banks (Figure 28 D)

Figure 28 An Example – Bank Loan Loss Provision Ratio



Sources: Bloomberg Finance LLP, FTSE Russell, Markit, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

There are a number of vendors providing industry-specific data:

- S&P Capital IQ: Compustat, Capital IQ Premium Financials, SNL
- Thomson Reuters (Worldscope, Reuters Fundamentals)
- Bloomberg

This is an active area of research and we expect to publish a number of papers focusing on each of main industries, e.g., banks, insurance companies, retailers, etc.

Country Specific Data

The dominant fundamental data vendors such as S&P Capital IQ and Worldscope try to reconcile company financial statements from different countries in one consistent manner. However, due to local customs, company fundamental data could be better supplied by specialized local vendors. A few examples include:

- China: WIND, DataYes
- Canada: Morningstar (formally CPMS)
- Japan: Toyo Keizai

Global Macroeconomic Data

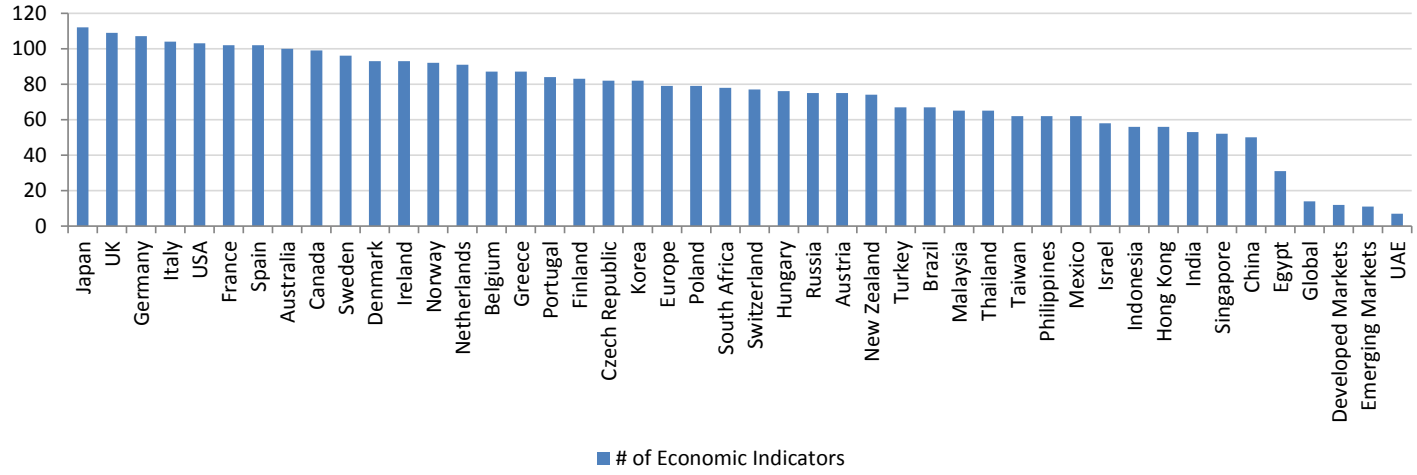
Economic data tends to suffer from significant reporting lags and restatement bias. Even simple questions such as how to retrieve the actual historical reporting date and time can be a huge challenge. In a series of forthcoming research papers, we will discuss our global macro data infrastructure in more details. Main data vendors include:

- Haver
- Thomson Reuter's Datastream
- S&P Capital IQ's IHS Global Macroeconomic Data
- Bloomberg
- FRED (Federal Reserve Economic Data)

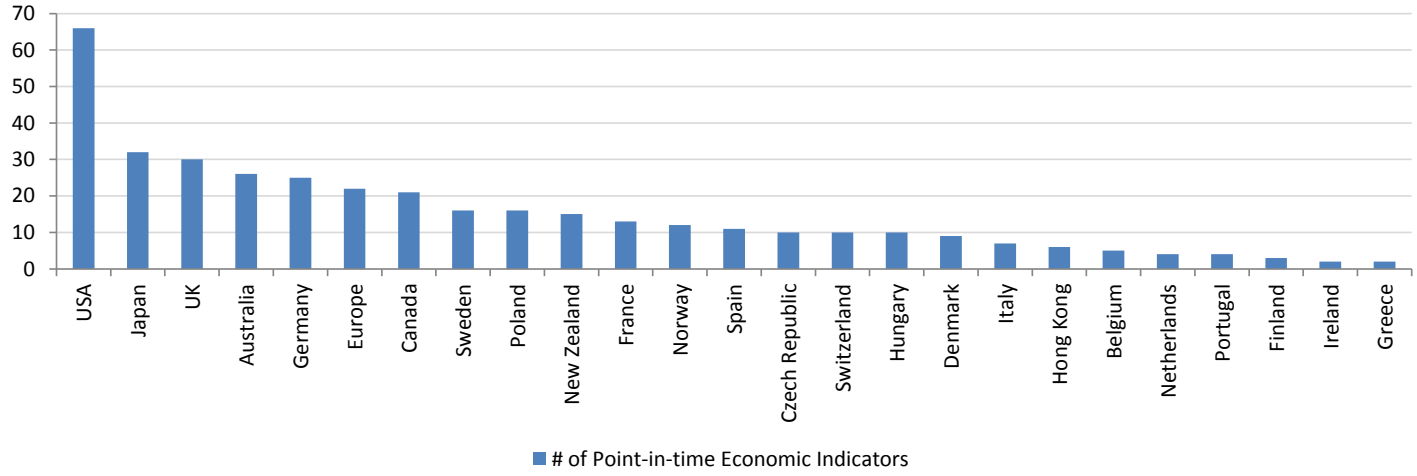
Figure 29 (A) shows the ~3,500 macroeconomic data series that we track for all countries and regions within the MSCI ACWI universe. Economic data in most countries are subject to regular revisions; therefore, it is also critical to have a true point-in-time economic database. As shown in Figure 29 (B), we also have ~400 point-in-time headline economic time series for most of the major countries globally.

Figure 29 Global Economic Database

A) Global Economic Indicators



B) Global Point-in-time Economic Indicators

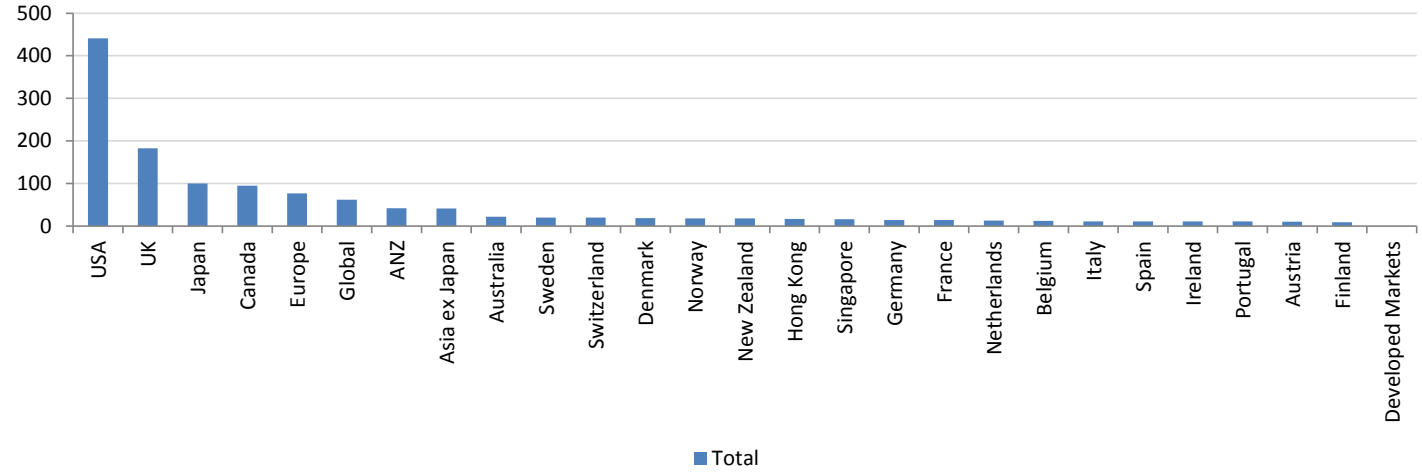


Sources: Bloomberg Finance LLP, FTSE Russell, Haver, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

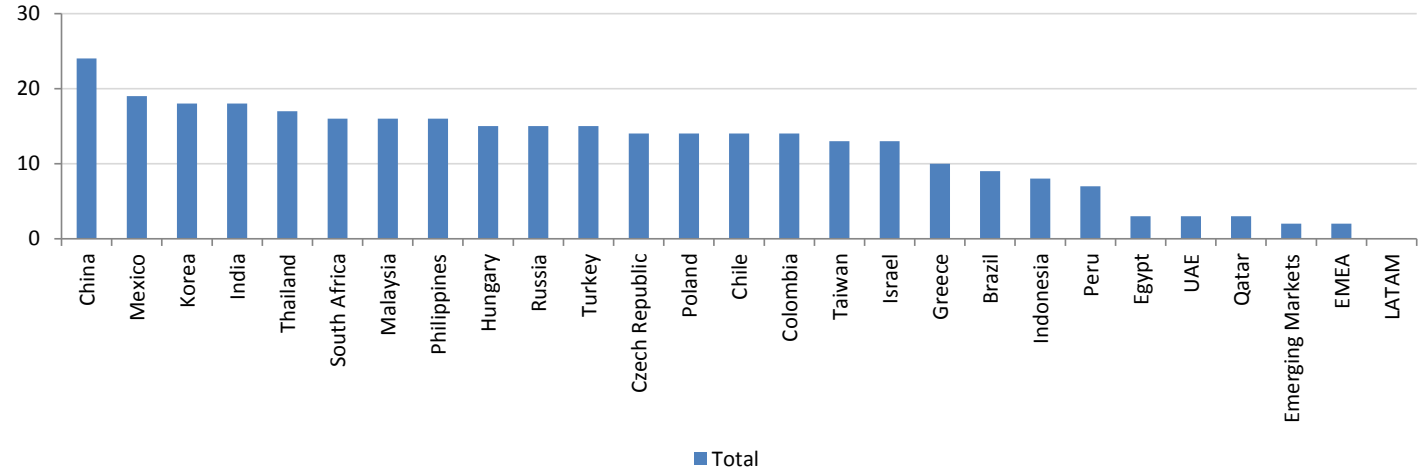
As shown in Figure 30, in addition to economic variables, there are also many other important macro data series to track, e.g., interest rates, sentiment, commodities prices, production, consumption, and inventories, etc.

Figure 30 Global Macro Database

A) # of Macro Indicators – Developed Markets



B) # of Macro Indicators – Emerging Markets



Sources: Bloomberg Finance LLP, FTSE Russell, Haver, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

Corporate Events, Transactions, Ownership, and People Data

There are hundreds of types of corporate events, from earnings announcements, executive changes, to conference presentations and new customer and product releases. A subset of more complex corporate events are related to transactions such as debt and equity offerings, M&As (Merger & Acquisitions), bankruptcies, and shareholder activism is being tracked closely by specialist event-driven hedge funds. For example, Figure 31 lists about 120 event types that are available from S&P Capital IQ's Key Development and Future Events database. In our previous research, we find event-based factors have strong alpha and tend to be uncorrelated to traditional factors.

Figure 31 List of Events from S&P Capital IQ's Key Development and Future Event Database

Main Category	Sub Category	Description	Main Category	Sub Category	Description
Financials	Guidance	Guidance - Lowered	Transaction	Buyback	Buyback Closings
Financials	Guidance	Guidance - Raised	Transaction	Buyback	Preferred Stock Buybacks
Financials	Guidance	Guidance - New/Confirmed	Transaction	Reorg	Business Reorganizations
Financials	Guidance	Guidance - Unusual Events	Transaction	Reorg	Strategic Alliances
Financials	Earnings	Earnings Announcement	Transaction	Financing	Seeking Financing/Partners
Financials	Earnings	Announcement of Interim Management Statement	Transaction	Financing	Shelf Registration Filings
Financials	Earnings	Announcement of Operating Results	Transaction	Financing	Follow-on Equity Offerings (Target)
Financials	Earnings	Sales Announcements	Transaction	Financing	End of Lock-Up Period
Financials	Customer/Product	Business Expansions	Transaction	Financing	Public Offering Lead Underwriter Change
Financials	Customer/Product	Client Announcements	Transaction	Financing	Private Placements
Financials	Customer/Product	Product-Related Announcements	Transaction	Financing	Debt Financing Related
Corporate Information	Corporate Structure	Address Changes	Transaction	Financing	Structured Products Offerings
Corporate Information	Corporate Structure	Changes in Company Bylaws	Transaction	Financing	Composite Units Offerings
Corporate Information	Corporate Structure	Fiscal Year End Changes	Transaction	Financing	Derivative/Other Instrument Offerings
Corporate Information	Corporate Structure	Legal Structure Changes	Transaction	Financing	Fixed Income Offerings
Corporate Information	Corporate Structure	Name Changes	Transaction	Spin-off	Spin-Off/Split-Off
Corporate Information	Corporate Structure	Exchange Changes	Transaction	Spin-off	IPOs
Corporate Information	Corporate Structure	Ticker Changes	Transaction	M&A	Considering Multiple Strategic Alternatives
Corporate Information	Corporate Structure	Delistings	Transaction	M&A	Potential Privatization of Government Entities
Corporate Information	Executive Change	Executive Changes - CEO/Chairman	Transaction	M&A	Seeking to Sell/Divest
Corporate Information	Executive Change	Executive Changes - CFO	Transaction	M&A	Seeking Acquisitions/Investments
Corporate Information	Executive Change	Executive/Board Changes - Other	Transaction	M&A	M&A Rumors and Discussions
Corporate Information	Index Change	S&P Events	Transaction	M&A	M&A Transaction Announcements
Corporate Information	Index Change	Index Constituent Adds	Transaction	M&A	M&A Transaction Cancellations
Corporate Information	Index Change	Index Constituent Drops	Transaction	M&A	M&A Transaction Closings
Corporate Information	Corporate Communication	Board Meeting	Legal	Auditor	Auditor Going Concern Doubts
Corporate Information	Corporate Communication	Annual General Meeting	Legal	Auditor	Auditor Changes
Corporate Information	Corporate Communication	Company Conference Presentations	Legal	Financials	Delayed Earnings Announcements
Corporate Information	Corporate Communication	Investor Conference	Legal	Financials	Discontinued Operations/Downsizings
Corporate Information	Corporate Communication	Earnings Calls	Legal	Financials	Impairments/Write Offs
Corporate Information	Corporate Communication	Operating Results Calls	Legal	Financials	Restatements of Operating Results
Corporate Information	Corporate Communication	Interim Management Statement Calls	Legal	Legal & Regulatory	Labor-related Announcements
Corporate Information	Corporate Communication	Fixed Income Calls	Legal	Legal & Regulatory	Lawsuits & Legal Issues
Corporate Information	Corporate Communication	Earnings Release Date	Legal	Legal & Regulatory	Delayed SEC Filings
Corporate Information	Corporate Communication	Operating Results Release Date	Legal	Legal & Regulatory	Regulatory Agency Inquiries
Corporate Information	Corporate Communication	Interim Management Statement Release Date	Legal	Legal & Regulatory	Regulatory Authority – Compliance
Corporate Information	Corporate Communication	Estimated Earnings Release Date (CIQ Derived)	Legal	Legal & Regulatory	Regulatory Authority – Enforcement Actions
Corporate Information	Corporate Communication	Sales Statement Release Date	Legal	Legal & Regulatory	Regulatory Authority – Regulations
Corporate Information	Corporate Communication	Guidance/Update Calls	Legal	Legal & Regulatory	Halt/Resume of Operations - Unusual Events
Corporate Information	Corporate Communication	Sales Statement Calls	Legal	Bankruptcy	Bankruptcy - Filing
Corporate Information	Corporate Communication	Shareholder/Analyst Calls	Legal	Bankruptcy	Bankruptcy - Other
Corporate Information	Corporate Communication	Special Shareholders Meeting	Legal	Bankruptcy	Bankruptcy – Asset Sale/Liquidation
Corporate Information	Corporate Communication	M&A Calls	Legal	Bankruptcy	Bankruptcy – Financing
Corporate Information	Corporate Communication	Special Calls	Legal	Bankruptcy	Bankruptcy – Reorganization
Corporate Information	Corporate Communication	Analyst/Investor Day	Legal	Bankruptcy	Bankruptcy - Emergence/Exit
Transaction	Dividend	Dividend Initiation	Legal	Bankruptcy	Bankruptcy - Conclusion
Transaction	Dividend	Dividend Increases	Legal	Bankruptcy	Debt Defaults
Transaction	Dividend	Special Dividend Announced	Legal	Credit Rating	Credit Rating - S&P - Upgrade
Transaction	Dividend	Dividend Affirmations	Legal	Credit Rating	Credit Rating - S&P - Downgrade
Transaction	Dividend	Dividend Cancellation	Legal	Credit Rating	Credit Rating - S&P - Not-Rated Action
Transaction	Dividend	Dividend Decreases	Legal	Credit Rating	Credit Rating - S&P - New Rating
Transaction	Dividend	Preferred Dividend	Legal	Credit Rating	Credit Rating - S&P - CreditWatch/Outlook Action
Transaction	Dividend	Stock Splits & Significant Stock Dividends	Activism	Activism	Investor Activism - Proposal Related
Transaction	Dividend	Stock Dividends (<5%)	Activism	Activism	Investor Activism - Activist Communication
Transaction	Dividend	Ex-Div Date (Regular)	Activism	Activism	Investor Activism - Target Communication
Transaction	Dividend	Ex-Div Date (Special)	Activism	Activism	Investor Activism - Proxy/Voting Related
Transaction	Dividend	Potential Buyback	Activism	Activism	Investor Activism - Agreement Related
Transaction	Buyback	Buyback Announcements	Activism	Activism	Investor Activism - Nomination Related
Transaction	Buyback	Buyback - Change in Plan Terms (Target)	Activism	Activism	Investor Activism - Financing Option from Activist
Transaction	Buyback	Buyback Tranche Update	Activism	Activism	Investor Activism - Supporting Statements
Transaction	Buyback	Buyback Cancellations	Activism	Activism	Activism

Sources: S&P Capital IQ, Wolfe Research Luo's QES

There are many vendors covering corporate events:

- S&P Capital IQ Key Development and Future Events
- S&P Capital IQ Transaction

- Thomson Reuters Transaction
- Thomson Reuters: Ownership, Activism
- Bloomberg
- 2iQ for insider transactions
- Factset: Deal Analytics
- News sentiment providers (Ravenpack, Thomson Reuters News Analytics, Wall Street Horizon, Alexandria)

While most vendors rely on analysts collecting and classifying events, computer algorithms can be designed to perform the task on a more timely fashion. Ravenpack[19], for example, has a suite of algorithms to classify a few thousands events.

Geospatial and Wireless Network Data

Satellite imagery data has gained acceptance in the investment community. Firms such as RSMetrics, SpaceKnow provide satellite imagery data. The most obvious application is to track traffic flows for retail companies (for example, at their warehouses and retail stores). The biggest challenge is often the actual location data for each and every company. Companies such as AirSage generate billions of anonymous location data points in the US, using wireless data from cellular phones.

- RSMetrics (Satellite Imagery)
- SpaceKnow (Satellite Imagery)
- AirSage (Wireless)
- EidoSearch (Pattern Recognition)
- Datascription (Text, audio, video)

Crowdsourcing

Crowd sourcing is the origin for many big data vendors. The information is collected through websites, surveys, or indirectly as a part of normal business activity of a company. While some vendors toil hard to source their data, Google has arguable the biggest platform for crowd sourcing. The Google Trends tool based on the Google search engine, provides search-volumes based trend analytics.

In addition to Google Trends, there are many other crowdsourcing data vendors:

- Estimize collects earnings and revenue forecasts for a large number of public companies beyond traditional sell-side analysts[20].
- Markit Securities Finance (formerly DataExplorer)

[19] We had the luxury to see a new beta product from Ravenpack. The company has significantly improved its event classification algorithm.

[20] In our previous research, we find earnings estimates from the Estimize database tends to be more accurate for large cap companies or when it is close to earnings reporting date, than conventional sell-side consensus.

- Quantopian offers a platform for anyone to conduct their own research and quantitative trading.
- Quandl provides clean data on financials, economic and alternative data to investors.
- LinkUp consolidates companies (both public and private) hiring data in real time.

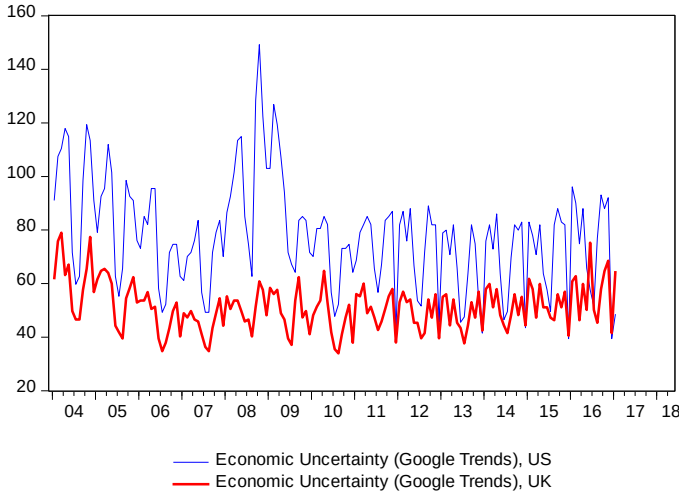
Google Trends and Tactical Asset Allocation

One of our favorite indicators to measure economic uncertainty is based on key word searches from Google Trends. As shown in Figure 32 (A), the Google Trends Economic Uncertainty indicator seems be at a record low in the US, but creeping up in the UK, possibly reflected the difference in views among the public in the US (in reaction to the new Trump administration) and UK (on Brexit). Because Google searches are done by both professional investors and the public, it shows a strong seasonality (see Figure 32 B). Obviously, there is much less interest on the economy when people are on vacation in the summer and December. Therefore, the seasonality adjusted Google Trends Economic Uncertainty in Figure 32 C is more revealing. Lastly, as shown in Figure 32 D, we can see that the measures of economic uncertainty by the public and by the professional investors are quite different[21]. Please note that we proxy the professional economic uncertainty by the cross-sectional dispersion of the real GPD predictions by a group of economists surveyed by Blue Chip Economic Indicator.

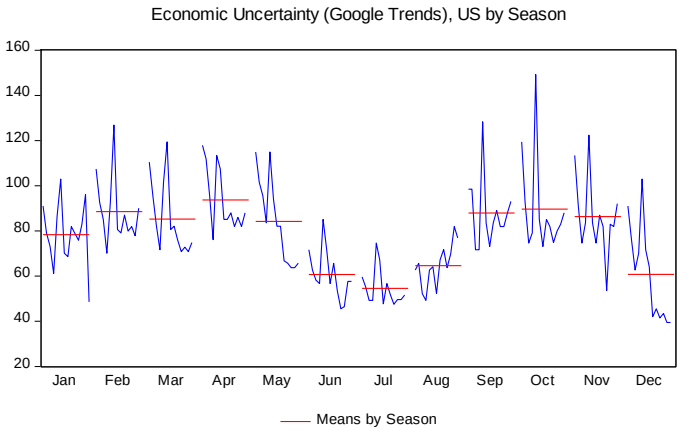
[21] The correlation between the two measures of uncertainty is less than 10%.

Figure 32 The Power of Google Trends

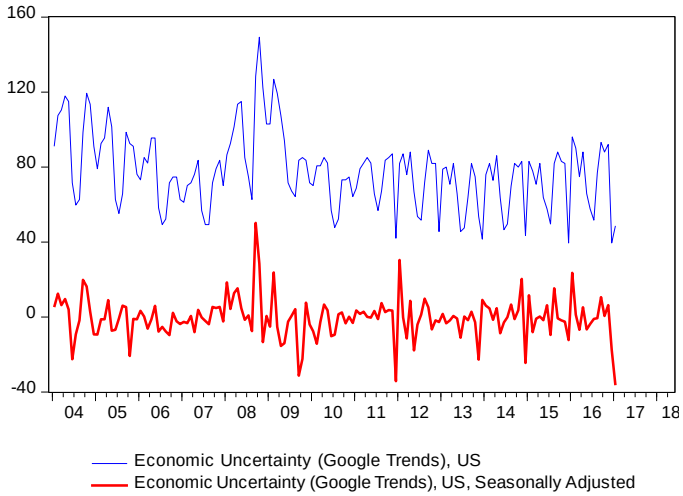
A) Google Trends Economic Uncertainty, US and UK



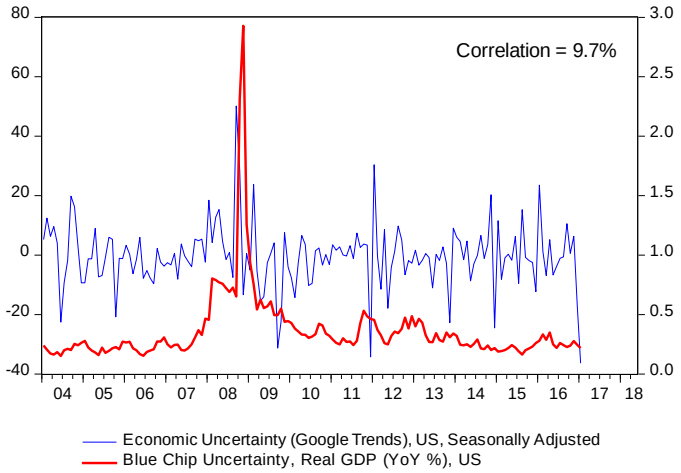
B) Seasonality



C) Seasonally Adjusted Google Trends Uncertainty



D) Google versus Professional Economists, US



Sources: Bloomberg Finance LLP, FTSE Russell, Markit, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

We now show an example of how Google Trends Economic Uncertainty can be useful in tactical asset allocation, by conducting the following time series regression:

$$r_t = \beta_0 + \beta_1 Uncertainty_{t-1} + \varepsilon_t$$

Where,

r_t is the return of an asset at time t, and

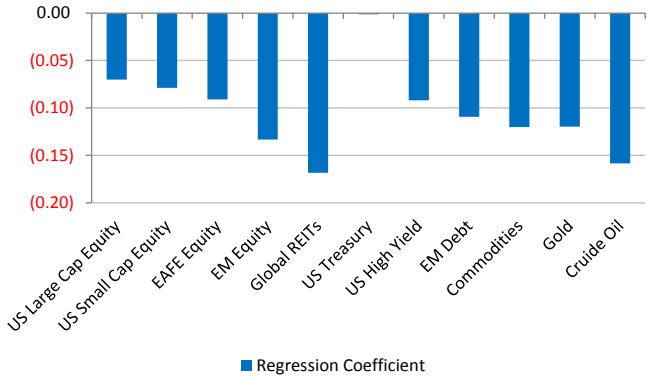
$Uncertainty_{t-1}$ is our Google Trends Economic Uncertainty at time $t-1$.

Figure 33 (A) plots the regression coefficient (β_1) for each of the 11 asset classes. It is interesting to note that the coefficients are all negative, suggesting a high economic uncertainty today leads to lower asset returns next month. The negative relationship is especially strong for REITs and crude oil.

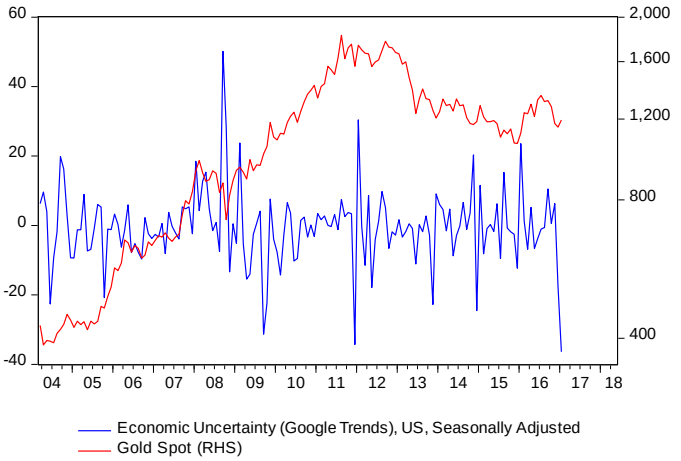
The relationships are not necessarily all intuitive, but the pattern appears to be strong (see Figure 33 B for an example of Gold price).

Figure 33 Google Trends Economic Uncertainty and Tactical Asset Allocation

A) Regression Coefficients with 11 Asset Classes



B) Google Trends Uncertainty versus Gold



Sources: Bloomberg Finance LLP, FTSE Russell, Markit, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

Crowdsourcing Stock Lending Data

Markit Securities Finance (formerly DataExplorer) provides a unique source of securities lending data. It collects information from a wide range of participants in the stock loan trading market, including beneficial owners, buy side investors, and intermediaries globally. It is important to note that the securities lending market is OTC; therefore, data on this market is scarce, incomplete, and delayed. Without crowdsourcing from various participants, it would otherwise be impossible to have a timely data on this market.

Traditionally, investors collect short interest data from exchanges (e.g., NASDAQ, NYSE, TSX in Canada) or regulatory filings (e.g., FINRA in the US and FCA in the UK). However, there are a number of issues:

- It is often reported with delays. For example, Compustat collects and consolidate short interest data for US and Canadian stocks, twice a month at mid-month and month end. The month end short interest reflects the short positions as of mid-month.
- It is not available in many markets.
- It only reflects the demand side, i.e., how many shares that short sellers want to short, but not the supply side, i.e., how many shares that asset owners are willing to lend.

Market Securities Finance database nicely fills in the gap. It has a number of interesting features to make it an invaluable source of information:

- It is updated daily, with a T+2 reporting lag[22].

[22] Data is also reported intra-day, with roughly 70% data available on a T+1 basis.

- It covers 30,000 equity instruments[23] globally.
- It has a large number of data fields, covering demand (short interest), supply (inventories available), cost of borrow, etc.

One simple yet interesting factor is to combine the demand and supply sides to measure the true short interest of a stock – the utilization factor:

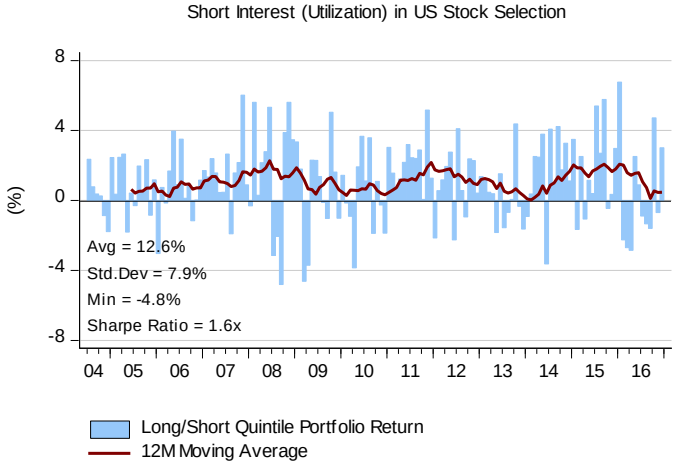
$$Utilization_{i,t} = \frac{ShortInterest_{i,t}}{AvailableInventory_{i,t}}$$

As shown in Figure 34 (A), a monthly rebalanced long/short portfolio based on the short interest (utilization) factor has delivered an average return of 12.6% and a Sharpe ratio of 1.6x, pre-cost[24].

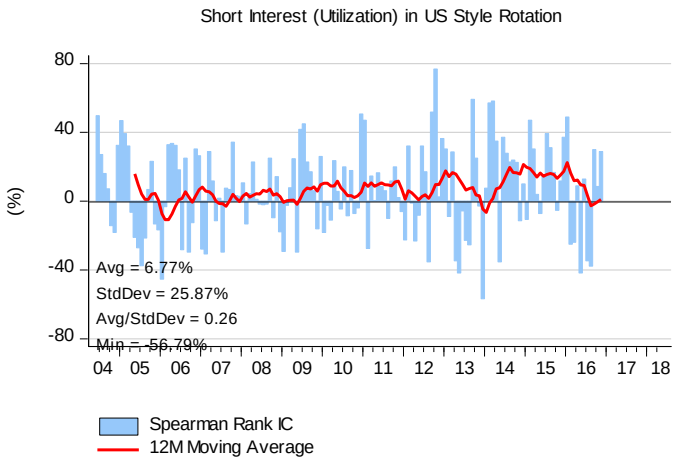
In addition to stock selection, the short interest (utilization) factor can also be used in a macro context. For example, Figure 34 (B) shows the performance of the signal in style (factor) timing. Each style factor is represented by a long/short equity portfolio. The backtesting on style rotation is based on 15 common factors. Performance is measured on Spearman rank IC, i.e., the correlation coefficient between the ranking of short interest (utilization) and forward one-month return of the style factors. Detailed methodology will be explained in a forthcoming research.

Figure 34 Markit Securities Finance – Short Interest (Utilization) Factor

A) Long/Short Quintile Equity Portfolio, US



B) Performance in US Style Rotation



Sources: Bloomberg Finance LLP, FTSE Russell, Markit, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

News, Text, and Sentiment

As investors would argue, the market and economic agents are not always rational. Sentiment plays an important role in asset pricing. We have a few ways to proxy investors' sentiment.

[23] Data is collected and reported on an issue level, including common shares, depository receipts, and ETFs. Markit also has short interest data on government and corporate bonds.

[24] The portfolio buys top 20% of stocks with the lowest short interest (utilization) and shorts the bottom 20% of stocks with the highest short interest. Stocks in both long and short sides are equally weighted. The performance is pre-transaction costs and assumes that we can short all stocks without stock lending fees. As will be discussed in a forthcoming research, stocks that are heavily shorted tend to be difficult to borrow and expensive to short. Therefore, the after-cost performance is likely to be different. We will elaborate how to best use the signal in the future.

The most traditional approach is to rely on sell-side research analysts. Investment banks and boutique research providers have a long history of supplying a large suite of metrics on the companies they cover, from EPS for the next quarter/year to buy-and-sell recommendations. The efficacy of analyst recommendation based signals has declined in recent years, but they remain a key part of investors' toolbox.

In recent years, quantifying news sentiment has gained recognition in the investment industry. Rather than relying on human investors to read and digest news and then come up with an investment thesis, NLP (Natural Language Processing) algorithms were developed to automatically quantify the sentiment in the textual information. Computer generated news sentiment tends to be fast (available in millisecond after the release of the news), unbiased, quantifiable, and cost efficient, but might not be as accurate as an experienced analyst. In our previous research[25], we find news sentiment is distinctively different from analyst sentiment, but the signals tend to be short-term in nature.

There are many news sentiment providers. Some of the well-known ones include: Ravenpack, Thomson Reuters News Analytics, Wall Street Horizon, Alexandria, NewsQuantified, Accern, Recorded Future, AlphaSense, and Benzinga.

Customer-Supplier Chain

Companies do not exist in isolation. They are connected to their customers, suppliers, competitors, joint ventures, and other related entities. The information contained in the customer-supplier chain provides a wealth of information on the operational and financial performance of a company. In our previous research[26], we find signals based on the supply chain data tend to have strong and uncorrelated alpha. Some of the best-known vendors include:

- Factset (Revere). Factset/Revere has one of the best coverage and probably the longest history on supply chain data. The data is more US centric, but does cover a large number of international companies.
- Bloomberg. Bloomberg probably has the most comprehensive coverage of supply chain data, but historical data is difficult to retrieve.
- Compustat. Traditionally, Compustat is the *de facto* source of quantifiable supply chain data for US companies.
- S&P Capital IQ Business Relationships
- Thomson Reuters Supply Chain

SRI/ESG/Forensic Accounting

Although SRI (Socially Responsible Investing) and ESG (Environmental, Social, and Governance) only gained momentum in recent years in the US, in many other parts of the world, such as Europe, Japan, and Australia, many pension funds require their external managers to be ESG-compliant to be considered for investing. For our purpose, we are more interested in whether ESG metrics also have additional alpha information. In our previous research, we find ESG-based signals have modest but

[25] We have studied using news sentiment data from both mainstream media and web blogs in global stock selection.
[26] We use supply chain data from Factset Revere in our research. We study models based on both statistical relationship among all related parties and pure customer-supplier chain data. We also extend the research using network algorithms (e.g., Google page rank).

uncorrelated alpha. More importantly, they are closely related to an exciting area of research on accounting quality. Some of the established vendors in this space include:

- MSCI
- Audit Analytics
- AAER
- Factset
- Thomson Reuters Asset4
- Bloomberg ESG (Bcause)

Multi-Asset Data

Portfolio managers used to be in silos. Equity managers only care about stock selection, while fixed income managers focus on the yield curve and credit spread. Data in other asset classes often contains untapped information. For example, we could tap into those hedge fund in the HFR (Hedge Fund Research) database, link to their holdings (sourced from Thomson Reuter's institutional ownership database) and use the "smart" holding signal to predict the returns of BioPharma stocks. There are many specialized hedge funds invested in BioPharma stocks; therefore, some of them are likely to have unique knowledge in this space. Similarly, we have studied stock-selection signals constructed from options data[27] and corporate bond data[28].

Some of the vendors providing multi-asset data include:

- Hedge Fund Research (HFR)
- OptionMetrics
- S&P Capital IQ Capital Structure
- Markit

Fund Flow Data

It has long been debated whether fund flow based factors are leading, concurrent, or lagging indicators.

Using US mutual fund flow data from ICI, we can show a simple example of how it can be used in factor timing. The signal is net new cash flow to active US equity mutual funds divided by total net

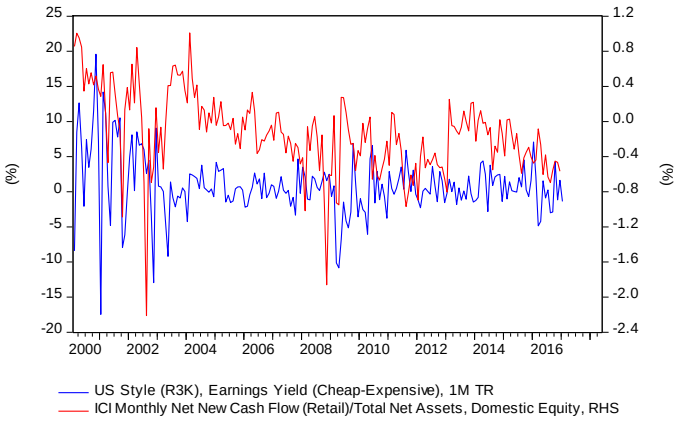
[27] There are a number of arguments why options data might lead equity returns. Options traders tend to be institutional in nature; therefore, are likely to be more sophisticated than the average stock investor. In addition, for investors with true insights and high conviction ideas, they may want to implement their strategies in the options market, due to the build-in leverage. We find even signals such as put-call parity (i.e., the average put-options-implied volatility – the average call-options-implied volatility) have strong predictive power of the returns of the underlying stocks, for up to a month. Separately, we also find an interesting macro indicator – we call it VRP or Variance Risk Premium – can strongly predict the returns of major asset classes for up to a year. The VRP is the difference between options-implied variance and realized variance. We will publish more on these topics in the near future.

[28] Most academic research suggests that the equity market tends to lead the bond market. However, in certain occasions, information from the fixed income market can be equally useful for equity investors. Credit analysts tend to fall into two camps – investment grade credit and high yield. For example, for high yield analysts, one of the biggest prizes is the issuers turn around their operations and the bonds are promoted to investment grade. Therefore, analysts covering distress debt might identify turnaround names better than equity analysts.

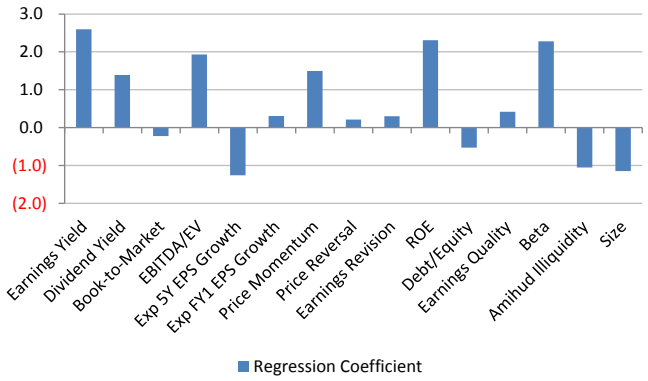
assets. Net new cash flow is new fund sales minus redemptions, combined with net exchanges. We want to understand the implications to equity styles when investors turn bullish on equities. As shown in Figure 35 (A), fund flow indicator appears to be positively correlated to value factor performance, with a lead of around three months[29]. This signal is predictive[30] to nine of the 15 most commonly used equity styles we track (see Figure 35 B). We will discuss style timing with more details in Part III (*Style Rotation, Machine Learning, and the New Frontier in Systematic Investing*) in this research series.

Figure 35 Fund Flow and Factor Performance

A) Fund Flow and Value Factor Performance, US



B) Fund Flow and Factor Performance, US



Sources: Bloomberg Finance LLP, FTSE Russell, Haver, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

Some of the vendors providing fund flow data include:

- EPFR
- ICI Mutual Funds and ETFs
- Morningstar
- Lipper

THE IMPORTANCE OF DOMAIN KNOWLEDGE

One of the biggest challenges of using alternative data sources (actually, for traditional databases too) lies in domain knowledge, i.e., a thorough understand of the specific subject, which may or may not be our expertise.

Expert knowledge about the specific field of study is critical to design potential factors, to understand whether the patterns identified in the backtesting make sense, to seek for alternative improvements, and to explain to our clients. In machine learning field, this is called feature engineering.

[29] A formal Granger causality test rejects the Null hypothesis that the fund flow indicator does not predict the value factor performance, with a p value of 3.5%.

[30] We assess the predictive power by performing the following regression: $r_{t+1} = \beta_0 + \beta_1 FundFlow_t + \varepsilon_t$. We use a Newey-West robust procedure to test whether the coefficient (β_1) is statistically significant.

Factor construction (or feature engineering) can be done either by analysts or automated via feature learning. We will explain this topic in more details in Parts II (*Signal Research and Multifactor Models*) and III (*Style Rotation, Machine Learning, and the New Frontier in Systematic Investing*) of this research series.

DATA SCIENCE FOR INVESTMENT MANAGEMENT

Once we have our technology infrastructure in place, before we can conduct investment analysis, however, there is one more extremely important step – data modeling. Traditionally, this was not a step that got much interest from researchers. However, as we also know, "once the data is the right shape of what we need, the job is 70% done". In recent years, there has been rapid development and interest in this field. Actually, even a new profession was born – the data scientist.

Building a fully automated trading system that makes money all the time has long been the dream to many investors – the Holy Grail of quantitative investing. With the rapid development in Big Data and sophisticated machine learning algorithms, it seems that we are getting closer to that dream. Nowadays, many large vendors (e.g., ClariFI, Factset, and Bloomberg) and increasing number of up-and-comers have launched tools to make quantitative backtesting easy. Off-the-shelf software often offers GUI (Graphical User Interface); therefore, with a point-and-click, you can backtest a large number of sophisticated factors and strategies. Systematic investing is no longer a game for a small group of elite mathematicians and finance professors. Suddenly, building a quant model seems to be so easy that almost anyone can have it set up with a modest upfront investment in technology.

However, the reality is far from that easy and rosy. As we have probably all witnessed, the live performance of a strategy is almost never the same as it is in backtesting. There are many intricate issues in data modeling that can cause serious biases or present us completely wrong results. For example, you may have already known what survivorship bias means, but you may not realize how much difference it can make. In this section, we discuss some of the well-known biases and propose solutions to some of the lesser-known problems.

SURVIVORSHIP BIAS

Survivorship bias is one of the most well-known, but interestingly, also one of the most common mistakes that investors tend to make. Most people are aware of the survivorship bias, but few understand its true significance. It is widely covered in academic literature, but few bother to quantify the real implications in backtesting. You will be surprised to see how many academic research papers and practitioners research still suffer from significant survivorship bias.

It is all too easy to backtest an investment strategy with the companies that are currently in the index. Tracking all companies that have ever existed in a correct point-in-time fashion, is actually not as easy as you may think. Many people would probably argue, if our interest is really on the companies that currently exist, maybe we should backtest a strategy using these companies anyway[31].

Companies come and go. A company can be delisted due to privatization, bankruptcy, acquisition, and prolonged period of underperformance. Similarly, new firms are formed and go to public. As shown in Figure 36 (A), the number of true point-in-time companies in the Russell 3000 index in the US stays relatively stable at around 3,000 over the past 30 years. However, among the 3,000 companies as of December 31, 1985, less than 500 have survived the years – there are only 428 remaining in the index now.

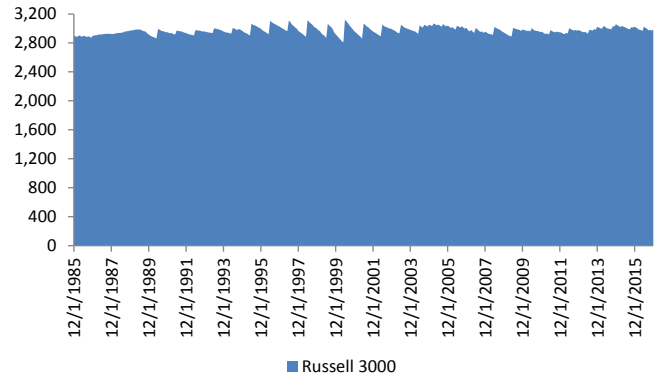
The phenomenon is even more pronounced in Europe. The BMI Europe Index, for example tracks the broad European market. It started with about 700 stocks in 1989 and now comprises almost 1,500

[31] We have seen a shockingly large number of analysts and data vendors pitching their products this way.

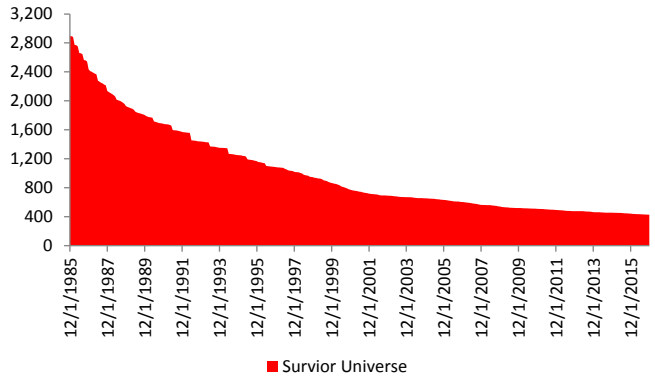
companies. Among the 700 stocks that were in the index at inception, less than 170 are still in the index today (see Figure 36 D).

Figure 36 # of Surviving Companies in the US and Europe

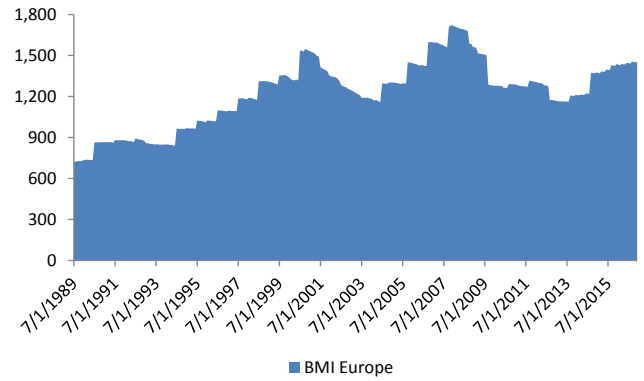
A) True Point-in-time # of Companies in the Russell 3000 Index



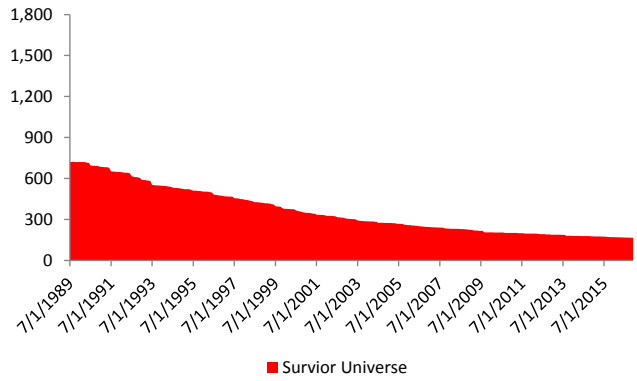
B) # of Surviving Companies in the Russell 3000 Index since 1985



C) True Point-in-time # of Companies in the BMI Europe Index



D) # of Surviving Companies in the BMI Europe Index since 1989



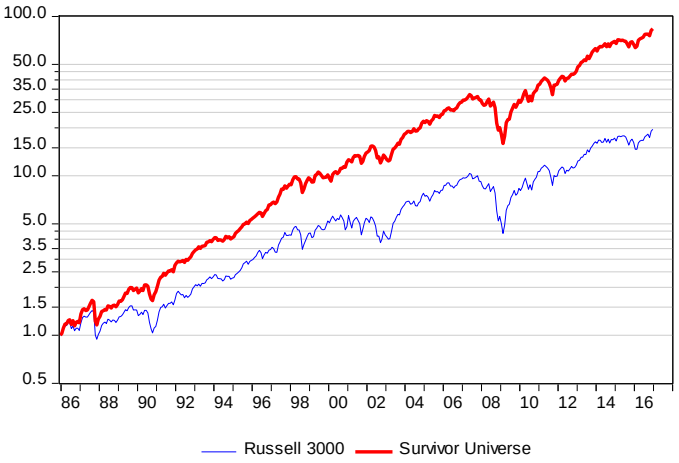
Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

Those companies that have disappeared tend to be the underperformers on average; and therefore, the survivors are the more successful ones. As shown in Figure 37, the survived companies have outperformed the broad market index by 427% and 217% in the US and Europe, respectively. The problem is, of course, that we do not know which company would survive in the future[32].

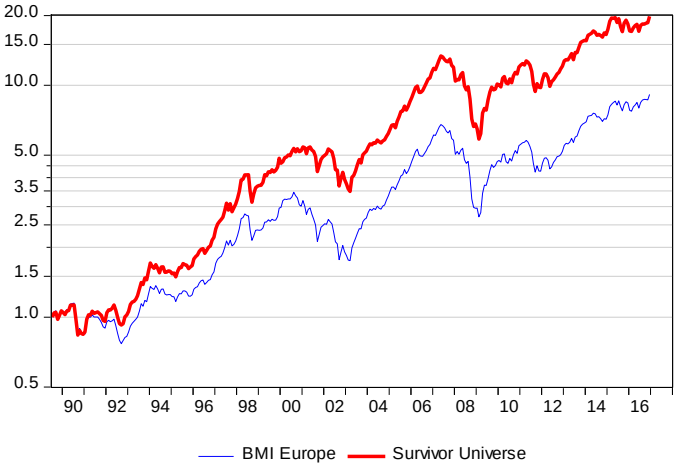
[32] There are many models and factors that can help us to assess distress risk and identify potential bankruptcy candidates, e.g., debt/equity ratio, Altman's z-score, Merton's distance to default. We will discuss these factors in Part II (*Signal Research and Multifactor Models*) of this series.

Figure 37 Survived Companies Outperform the Broad Market

A) US



B) Europe



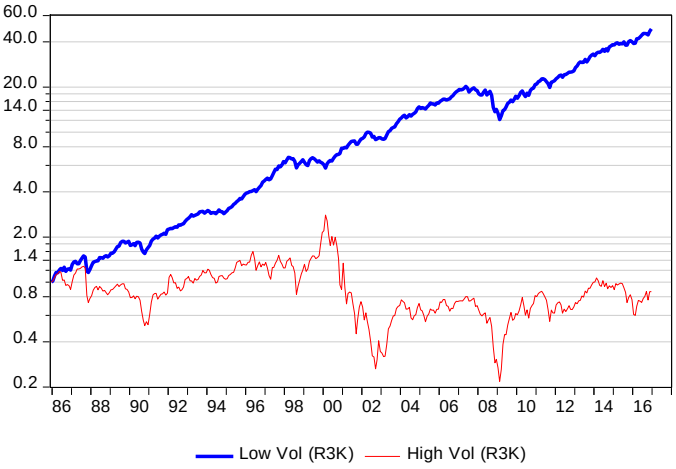
Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

Backtesting with only those survivors creates a considerable bias and in many occasions, produces completely wrong conclusions. Now, let's prove this argument with one of the most commonly used "backtests" by many practitioners. Because tracking companies point-in-time over a long history is not easy, many practitioners and data vendors conduct backtesting using the current index constituents. They contest, since you can only invest in the companies existing today, there is nothing wrong to do research only on these firms. The problem is, back in time, we would not know which companies could survive in the future and what firms could be created or successful enough to be added to the index down the road. Backtesting with today's index constituents can create completely wrong investment conclusion.

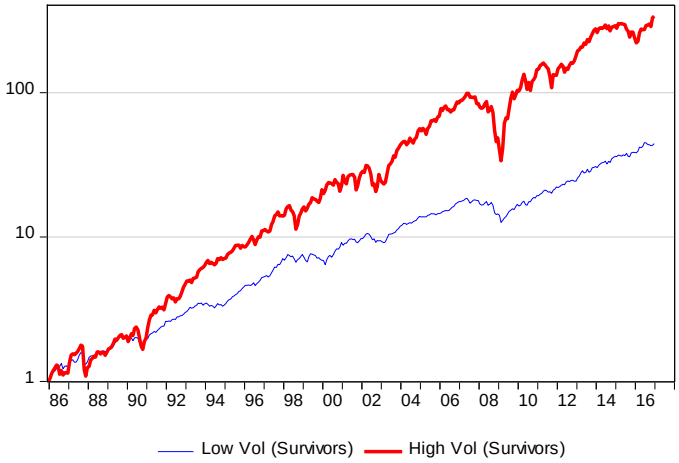
Let's start from a simple factor – the low volatility anomaly, which argues that stocks with lower volatilities tend to outperform high risk stocks in the long term. A proper backtesting using the point-in-time Russell 3000 universe finds that low vol stocks outperform high vol stocks significantly over the past 30 years (see Figure 38 A). However, if we repeat the backtesting, but use those companies that have survived until today, the result is exactly the opposite – now high vol stocks beat low vol stocks by 7.6x (see Figure 38 B). The seemingly surprising result is actually intuitive. Many volatile companies have gone out of business, but the ones that did survive are the turnaround stories with considerable upside (only if we have the foresight of which firm would survive).

Figure 38 The Low Vol Anomaly

A) Using a Point-in-time Universe



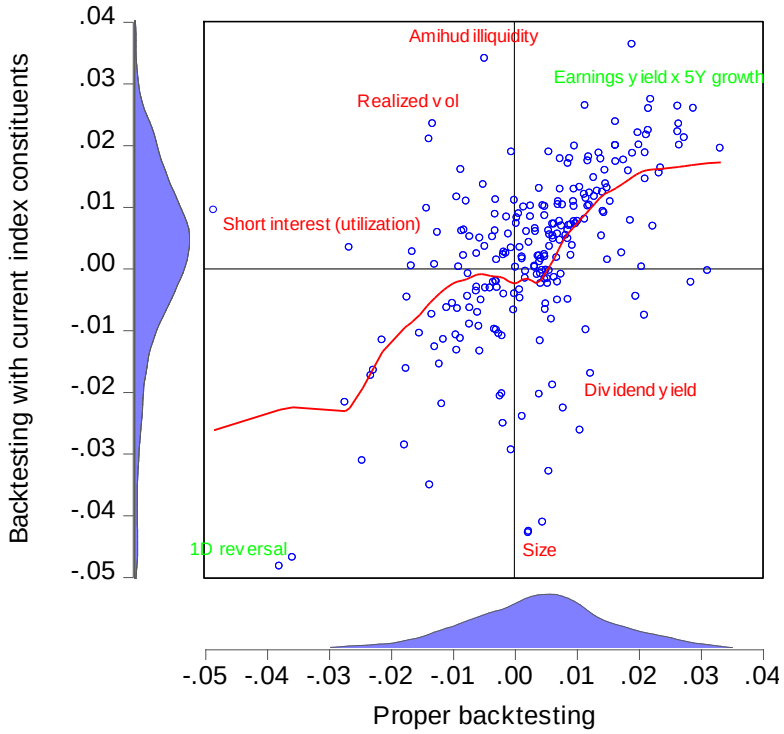
B) Using the Survived Companies



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

The difference does not stop here. Now, let's conduct a broader experiment of 235 common stock-selection factors. We perform two backtestings – one using a proper point-in-time universe of the S&P 500 index (i.e., the actual companies in the S&P 500 index since 1985, at each month end) and one using only the 500 companies in the index today. It is shocking to see, not only the difference between these two backtestings is significant (i.e., greater than 10%) for the vast majority of factors (>90%), but also have completely opposite signs for over 30% of the factors. Figure 39 shows the difference between these two backtestings – any factors fall into the upper-left and bottom-right quadrants have opposite signs based on the two approaches. For example, using a proper point-in-time universe, stocks with higher dividend yields outperform those non-dividend paying companies. However, if you use today's 500 companies in the index, you would find the opposite. Figure 40 summarizes the top and bottom 15 factors with largest difference.

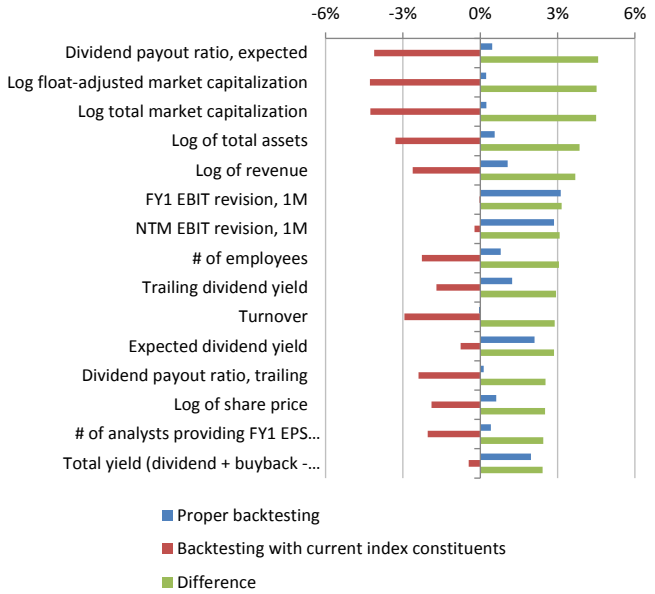
Figure 39 The Difference between Proper and Survivorship-biased Backtesting



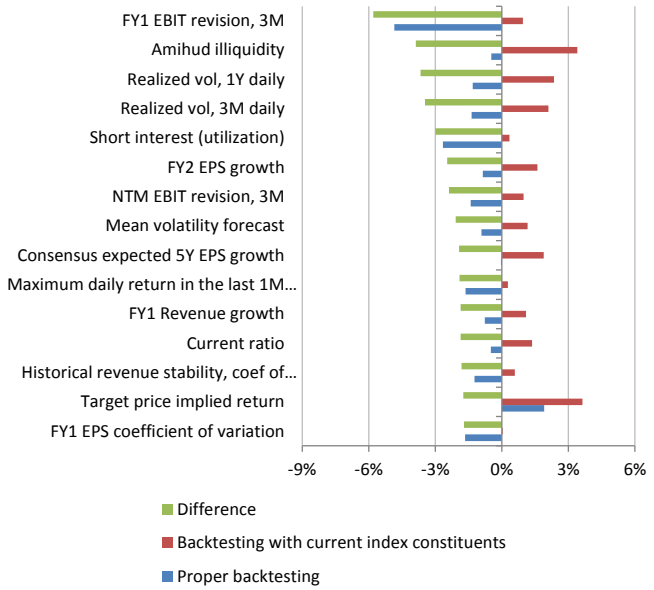
Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

Figure 40 Factors with the Largest Differences

A) Top 15 Factors



B) Bottom 15 Factors



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

LOOK-AHEAD BIAS AND POINT-IN-TIME DATA

The second and even more damaging problem is the look-ahead bias. It is the bias created by using information that was unknown or unavailable as of the time when the backtesting was conducted. It is the most common mistake that we have seen. The survivorship bias can be considered as a special case of the look-ahead bias, because the question whether a stock will survive (or be added in the index) in the future is unknown on the backtesting date.

More and more data vendors have realized the importance of point-in-time databases. Point-in-time means the exact information that was available to the market participants (or, in the case of vendors, the exact data shown in a vendor's database) as of a given point of time. Point-in-time database allows analysts to use the "freshest" data as of any given point of time to construct the most realistic investment strategies.

Without point-in-time databases, researchers had to rely on the traditional databases. The problems with the traditional databases are three fold.

- First, analysts have to make **reporting lag assumptions**. For example, we can't use quarterly EPS for calendar quarter ending December 31, 2000 on January 31, 2001 for all companies, because some companies had not reported their Q4/2000 earnings yet. Therefore, analysts would typically add a few months of reporting lag. However, this process also adds stale information. Using our previous example, as of January 31, 2001, many companies, especially, large cap companies had already reported the calendar quarter ended on December 31, 2000. Assuming all companies have two months of reporting lag forces us not to use the information actually available to us. Point-in-time data effectively solves the reporting lag assumption problem – we do not need to make such an assumption at all. We simply use the best information available as of any given point of time for our research. In this case, if a company has reported Q4/2000 results, we would use Q4/2000 data; otherwise, we would use Q3/2000 (or whatever was available as of January 31, 2001).
- Companies often **re-state** their financial statements due to corrections to accounting errors or changes in accounting policies. Traditional databases only keep the latest numbers or the last re-stated financial statements. For analysts trying to build realistic investment scenarios going back in time, analysts would be using information that was actually not available as of the backtesting period. Another form of look-ahead of bias arises when data vendors add new companies. Periodically, vendors add new companies to their databases. When they do so, they often add a few years of historical financial statements in the system too. If we use the current databases, we are using companies that were not actually in the databases in the backtesting period, which often introduces overly optimistic results.
- The third problem in the traditional databases is **survivorship bias**. As companies engaged in M&A, bankruptcy, delisting, and other forms of corporate actions, companies and stocks are being removed from the database constantly.

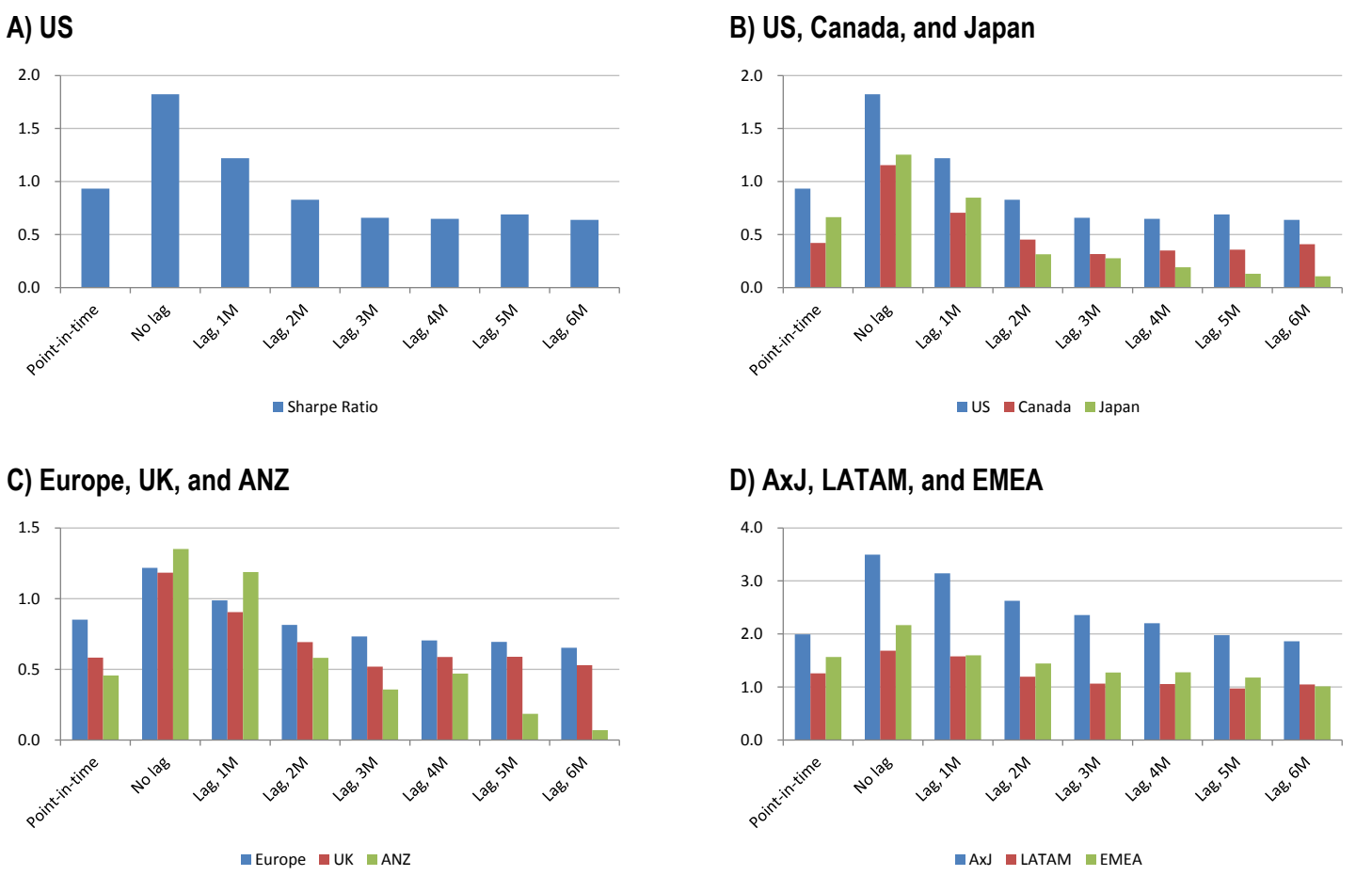
The Impact of Reporting Lag Assumptions

To show the impact of look-ahead bias and reporting lag assumption, we conduct a monthly backtesting using a value factor (trailing earnings yield – EPS/Price). The benchmark is a proper point-in-time database without look-ahead bias, using the actual EPS data as of each month end. Then we compare the performance with a backtesting with full look-ahead bias, using the

corresponding EPS data without any reporting lag, i.e., assuming EPS data become available immediately after the close of fiscal quarter (or any other reporting period). Lastly, we add a series of reporting lags, from one to six months.

As shown in Figure 41 (A), look-ahead bias inflates the performance of our value factor by almost 100% in the US! The impact of look-ahead bias is evident in all regions. In the US, Canada, and Japan, it appears that a reporting lag between one and two months produces the most consistent result as the proper point-in-time data (see Figure 41 B). Adding a reporting lag beyond two months introduces too much stale information and drags down the performance significantly. In Europe, UK, and ANZ, a lag assumption between two and three months is appropriate (see Figure 41 C), while for AxJ, LATAM, and emerging EMEA, we need to increase the lag assumption to three months (see Figure 41 D). The different lag assumptions reflect how timely companies in each region report their earnings.

Figure 41 The Impact of Reporting Lag Assumptions around the World



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

Other Forms of Look-Ahead Bias

Not all forms of look-ahead biases are so obvious. One of the most common questions that we hear from our corporate clients is on the optimal level of share price to target. Most companies probably do

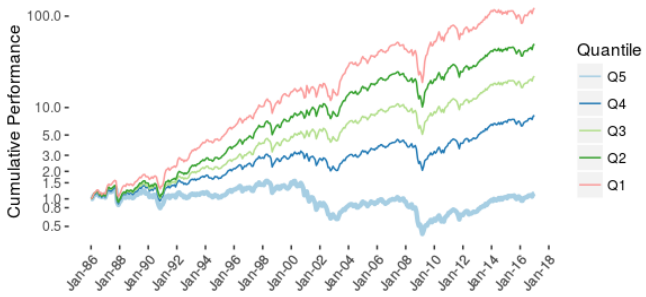
not want their share price to be above $1,000, which would prohibit many retail investors from owning their stocks. Therefore, successful companies that see their stock prices rising too high would divide each share of stock into a multiple of one share, i.e., stock split. Similarly, since penny stocks are generally perceived as very risky, once a company's share price drops below $5, it could apply a reverse-split to shore up its price.

In theory, share price itself should be irrelevant to a firm's valuation. The total value of a company is determined by the market. Share price is only a function of the number of shares outstanding. In practice, the level of share price may attract a special kind of client, i.e., the so-called clientele effect. For example, low priced stocks are more likely to induce speculators.

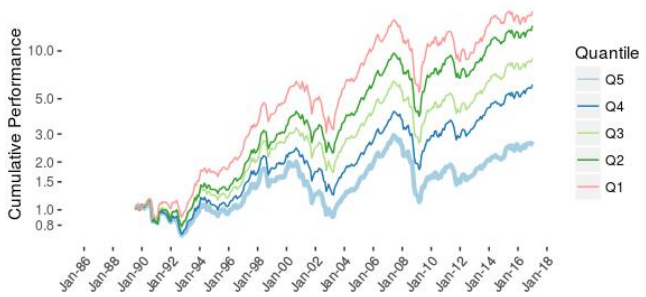
A simple backtesting of the share price factor – yes, the factor is defined as the share price itself – turns out to be surprising. As shown in Figure 42 (A) and (B), low priced stocks have consistently outperformed high priced companies in both the US and Europe in the past 30 years. The story telling nature would say, "this is a classic example of taking on risk (risk of penny stocks) and then getting compensated for that risk". However, before we start to celebrate the finding, we have to admit that it is in conflict with our prior view on stock splits (see Grinblatt, Masulis, and Titman [1984]). In our event study, we observe abnormal positive returns on the announcement day and positive post announcement drift for companies that split their stocks (and negative returns for reverse splits). If outperforming stocks are more likely to split and splits further boost share price, why high priced stocks underperform penny stocks?

Figure 42 The Performance of the Share Price Factor – Using Split-Adjusted Price

A) US



B) Europe



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

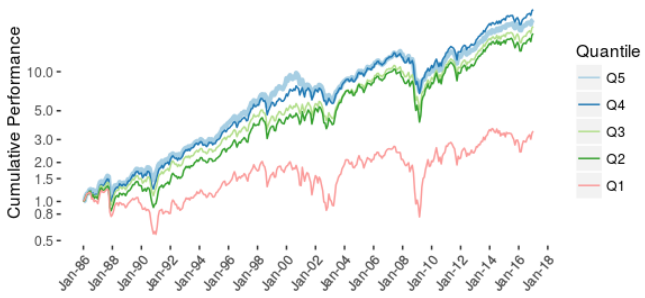
We need to be particularly careful with stock splits. Most data items that we get from most vendors are split adjusted, which is mostly what we need. For the vast majority of factors, we only need split adjusted data. For example, when we compute price-to-earnings multiple, we want both price and EPS split adjusted, as long as the adjustments are consistent. As price data is from our market data database, while EPS comes from our fundamental database, the split adjustment may not always synchronize. Similarly, when we plot a chart on Bloomberg showing the price of a stock, we want to see split-adjusted price.

However, when we deal with factors that rely on only price, trading volume data, or EPS, we need to be extra careful. The backtesting shown in Figure 42 is based on split-adjusted prices. Many investors either are not aware of split-adjustments or do not believe it makes such a big difference – aren't stock splits rare events?

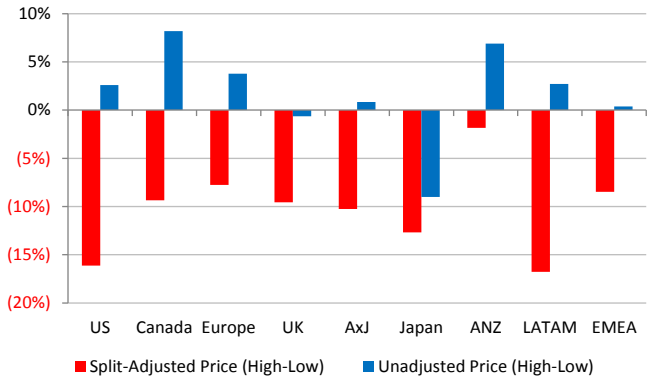
If we repeat our backtesting with the correct unadjusted price, as shown in Figure 43 (A), now the results completely flip to the opposite direction. Low priced stocks significantly underperform in the US. Outside of the US, we witness exactly the same pattern in almost all regions, with the exception of Japan. Backtesting using split-adjusted price suggests high priced stocks lag behind, while correct unadjusted price presents the opposite (see Figure 43 B).

Figure 43 The Performance of the Share Price Factor – Using Un-Adjusted Price

A) US



B) Global Evidence

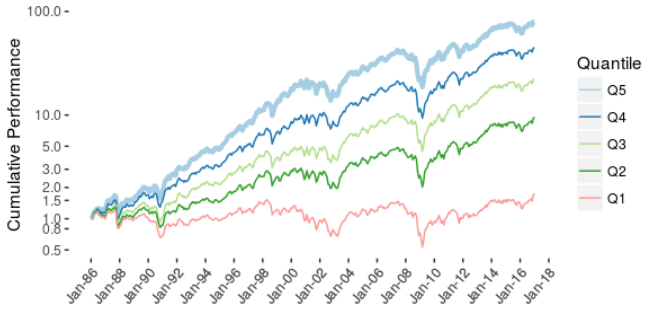


Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

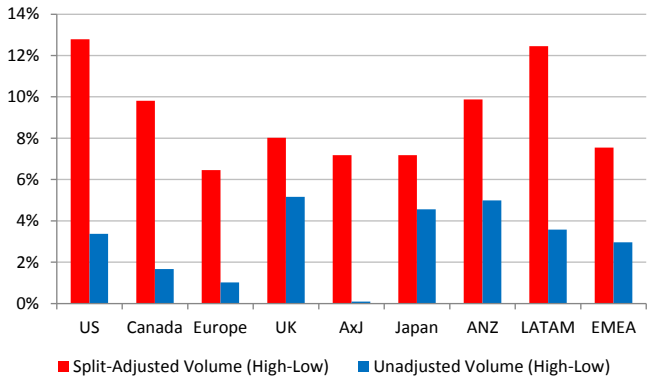
The corporate action bias is not limited to share price. For example, it also impacts trading volume. The backtesting on trading volume mirrors what we find for share price. As shown in Figure 44 (A), using split-adjusted volume data infers stocks with high trading volumes offer significantly higher returns. However, that is largely an artifact. Outperforming stocks tend to have higher share prices with multiple splits; therefore, split-adjusted volume overstates the true volume at the time back in history. The proper backtesting using unadjusted trading volume shows a very different picture – now trading volume is no longer a good stock-selection signal (see Figure 44 B).

Figure 44 The Performance of the Trading Volume Factor

A) Split-Adjusted Trading Volume Factor in the US



B) Global Evidence



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

DATA PRE-PROCESSING

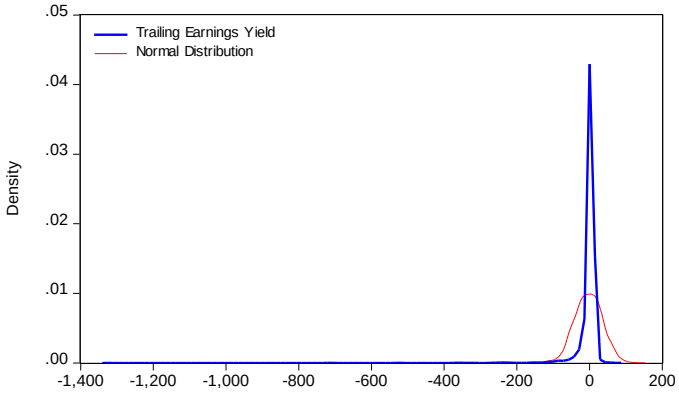
Data pre-processing refers to the process of checking for potential errors and outliers, removing irrelevant and erroneous data points, and transforming raw data. In machine learning literature, data pre-processing is also known as *feature engineering*. As we will discuss later, different modeling techniques have different sensitivity to data issues. For example, if we measure factor performance using Spearman rank correlation, the impact of outliers is minimal. On the other hand, the Pearson's correlation is highly sensitive to outliers. Similarly, linear regression type of models is sensitive to data transformation, while CART (classification and regression tree) is robust to skewed data. How the input data is encoded can have a significant impact on model performance.

Data Distribution, Skewness, Kurtosis, and Outliers

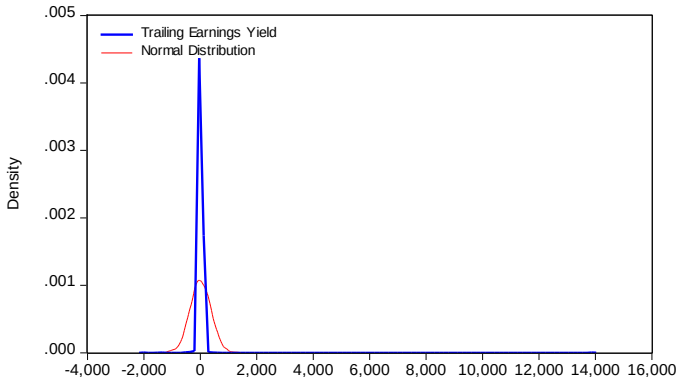
Extreme outliers are common in financial data. Figure 45 (A) and (B) plot the distribution of value (trailing earnings yield) factor in the US and Europe, as of December 31, 2016. It is evident that the distribution deviates greatly from the normal distribution in both regions.

Figure 45 Factor Distribution

A) Value Factor Distribution, US



B) Value Factor Distribution, Europe



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

Data Transformation

In addition to outliers, financial data is also often highly skewed with excess kurtosis. For example, the value factor is more skewed to the negative side in the US and positive side in Europe (see Figure 45). Statistical models often encounter issues with such data. Therefore, it is generally advised to transform your raw data to as close to normal distribution as possible, without too much distortion.

Z-score transformation

Z-score transformation is the most commonly used data normalization technique. It essentially centers the raw data to have zero mean and scales the data, so the result has a standard deviation of one:

$$f_z = \frac{f - @mean(f)}{@stdev(f)}$$

The z-score transformation centers and scales the data with a sample mean of 0 and a standard deviation of 1. However, it does not make any adjustment to skewness and kurtosis. As shown in Figure 46 (A), the z-score factor still has negative skewness and excess kurtosis. Furthermore, outliers still exist in the same way.

Winsorization versus Truncation

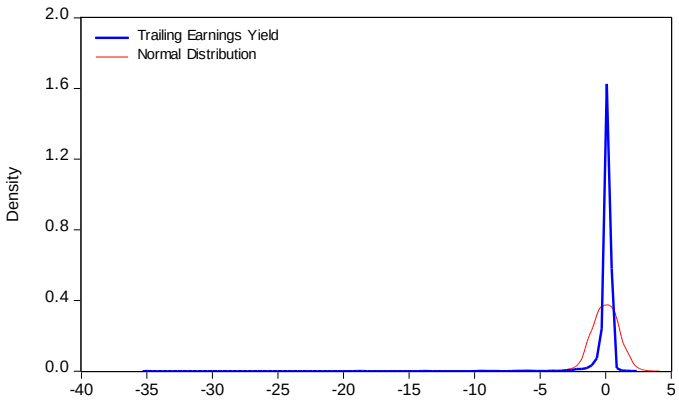
More often than not, outliers are caused by data errors or rare events that are unlikely to be repeated in the future. Therefore, there is little useful information on outliers. In practice, outliers can be removed from our sample completely – a technique called trimming, truncation, or censoring in econometrics. Figure 46 (B) shows the distribution after a truncation is applied, where we remove all samples greater than ± 3 standard deviation. After the censoring, the negative skewness/excess kurtosis problem is lessened significantly, but the data is still far from a standard normal distribution. The most damaging effect of truncation is that it shrinks our sample. As we remove these observations from our data, we can't express any views on these stocks. Therefore, the alternative to truncation, the winsorization is more commonly used.

Instead of truncation, we can "winsorize" data by replacing outliers that are greater or less than certain pre-defined xth percentile (e.g., 1%, 5%, or 10%) with the xth percentile values. Winsorization and truncation seem to produce similar results. However, comparing (C) with (B) in Figure 46, we can see the small spike of observations at -3. For observations with less than -3, winsorization keeps the value at -3; while truncation removes these samples all together.

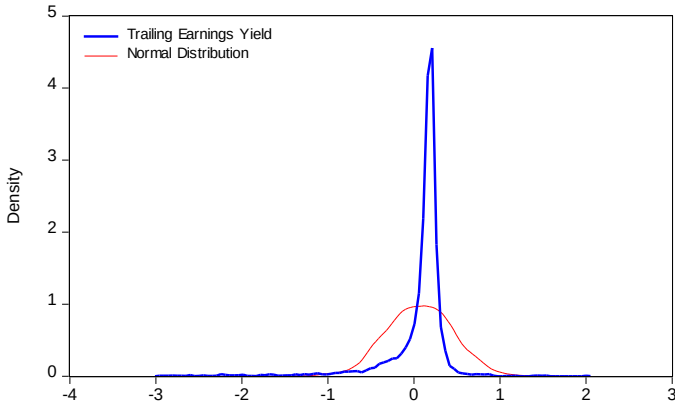
We can further improve the distribution by repeating the z-scoring/winsorization a few times (see Figure 46 D).

Figure 46 Z-Score Transformation

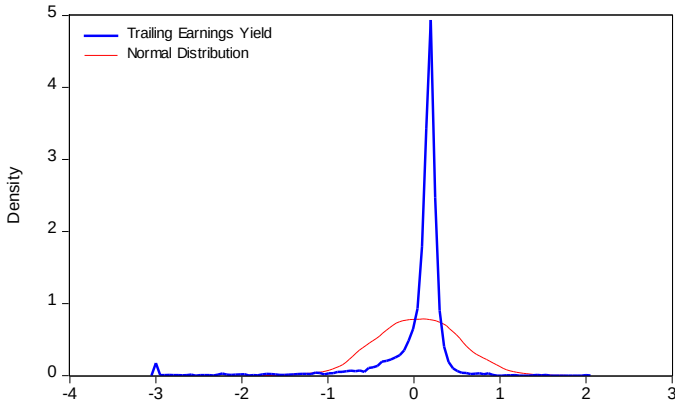
A) Z-Score Transformation of a Value Factor, US



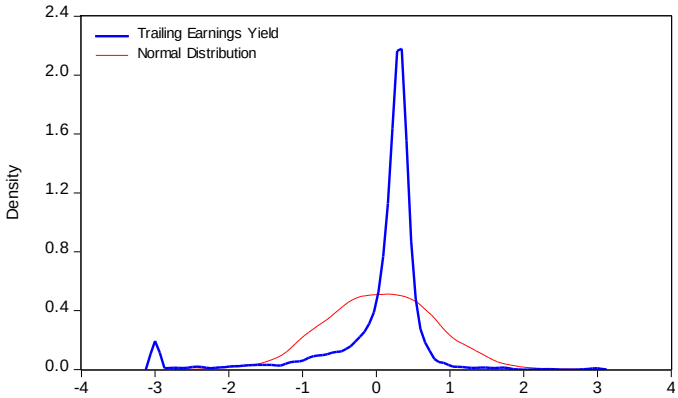
B) Z-Score Transformation with Truncation



C) Z-Score Transformation with Winsorization



D) Z-Score Transformation with Repeated Winsorization



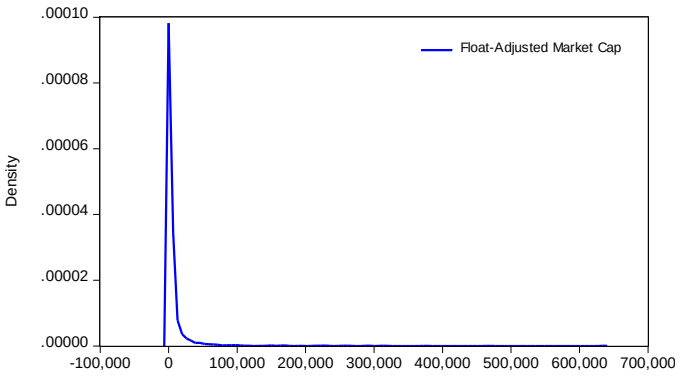
Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

Box-Cox Transformation

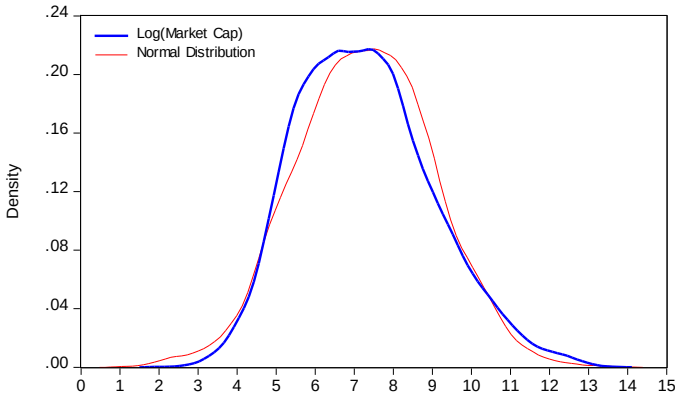
Raw data can also be highly skewed, which may cause issues for some models. To adjust highly skewed data, we could use logarithm, square root, or inverse transformation. As shown in Figure 47 (A), the original market cap distribution is highly skewed, as market cap is bounded above 0. The logarithm transformation is effective to transform the data to a nearly perfect normal distribution (see Figure 47 B).

Figure 47 Market Cap Distribution

A) Market Capitalization Distribution, US



B) Log Transformation



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

Box and Cox [1964] propose a family of transformations that are indexed by a parameter, λ:

$$f_{BC} = \begin{cases} \frac{f^{\lambda} - 1}{\lambda} & if\ \lambda \neq 0 \\ \log(f) & if\ \lambda = 0 \end{cases}$$

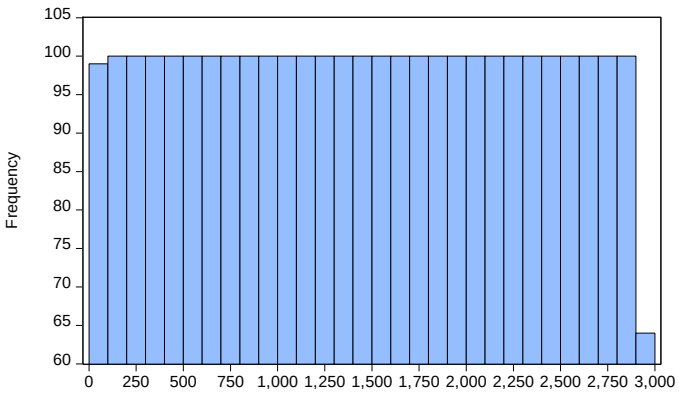
In fact, logarithm, square root and inverse transformations are all special cases of Box-Cox transformation.

Ranking, Percentile, and Quantile Transformation

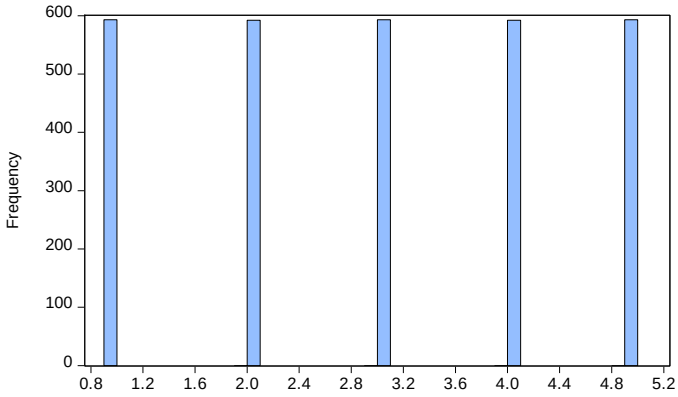
Another way to transform data is the ranking, percentile, and quantile approach. We essentially argue that all that matters is the ranks, while the actual distance in factor scores is uninformative. This is also our preferred and default data transformation algorithm. The ranking transformation automatically converts data (of any distribution) to a uniform distribution. All outliers are pushed to the normal range (see Figure 50 A). It is also common to bucket observations into percentile or other quantiles (e.g., deciles, quintiles, quartiles, or terciles). However, as the number of buckets becomes smaller, we lose more and more information (see Figure 50 B).

Figure 48 Rank, Percentile, and Quantile Transformation

A) Rank Transformation, Value Factor (US)



B) Quintile Transformation, Value Factor (US)



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

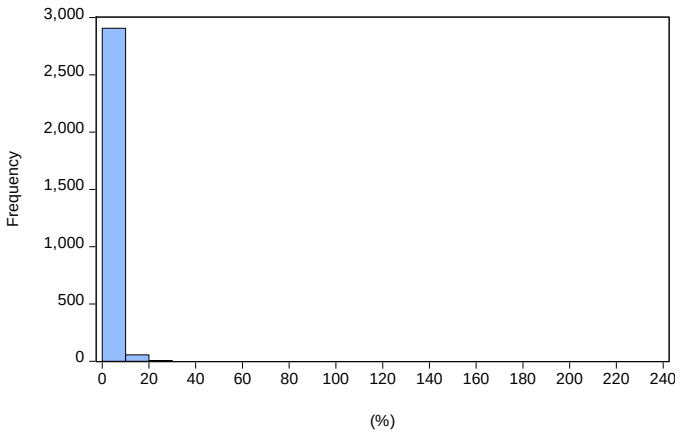
Discrete and Sparse Values

The other issue that arises from data transformation is that the distribution of factor scores can be in discrete values. For example, the Piotroski's F-Score (see Piotroski [2002]) and Mohanram's G-Score (see Mohanram [2005]) are both based on summing up a suite of binary indicators. The Piotroski's F-Score takes on a value between 0 and 9, while the Mohanram's G-Score is a discrete number from 0 to 8. We will have a problem to form decile portfolios with either factor.

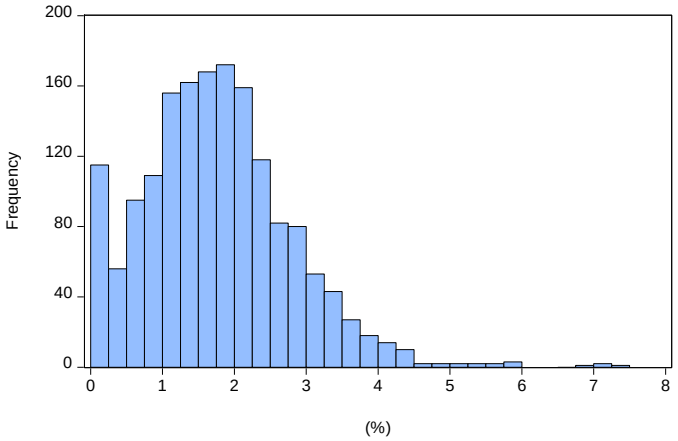
A more intricate case is, when the factor itself does not appear to be discrete, but the actual values are concentrated in a given number. For example, currently around 46% US companies do not pay any cash dividend; therefore, their dividend yields are all zero. As shown in Figure 49 (A), the distribution of the dividend yield factor is dominated by those zero values, which causes issues for both quantile portfolio and IC calculations. On the other hand, almost all companies in Japan pay regular dividends; therefore, the distribution of dividend yield factor has a much better shape in Japan (see Figure 49 B).

Figure 49 Discrete and Sparse Values in Factor Distribution

A) Dividend Yield Factor in the US



B) Dividend Yield Factor in Japan



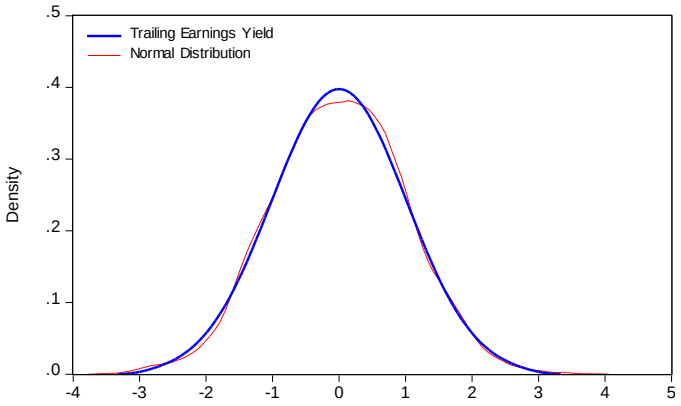
Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

Our Proprietary Data Transformation Technique

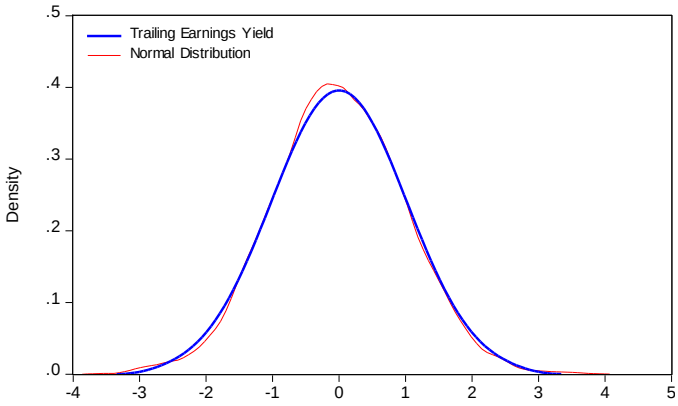
From ranking, we further transform the data into a standard normal distribution, by applying an inverse normal mapping. As shown in Figure 50 (A) and (B), our proprietary algorithm successfully transforms most factors to an almost perfect standard normal distribution

Figure 50 QES Proprietary Data Transformation

A) Value Factor in the US



B) Value Factor in Europe



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

Does it Matter?

We seem to have spent quite a bit of time to discuss data transformation. You may start to wonder how much difference data transformation makes. Most analysts probably think the impact is minimal, but you may well be surprised.

We show the impact of data normalization with a simple eight-factor model:

- Value: Trailing earnings yield – we prefer companies with high earnings yield

- Value: Book-to-market – we buy companies with high book-to-market, i.e., cheap stocks on valuation
- Growth: Historical year-over-year interim EPS growth – we prefer companies with high earnings growth
- Price Momentum: 12M total return – we prefer companies with positive price momentum
- Analyst Sentiment: 3M EPS revision – we buy companies with positive earnings revisions
- Quality – Profitability: Return on equity – we like firms with high ROEs
- Quality – Leverage: Debt/Equity ratio – we prefer companies with low financial leverage
- Quality – Earnings Quality: Sloan's accruals (see Sloan [1996]) – we buy companies with low accruals

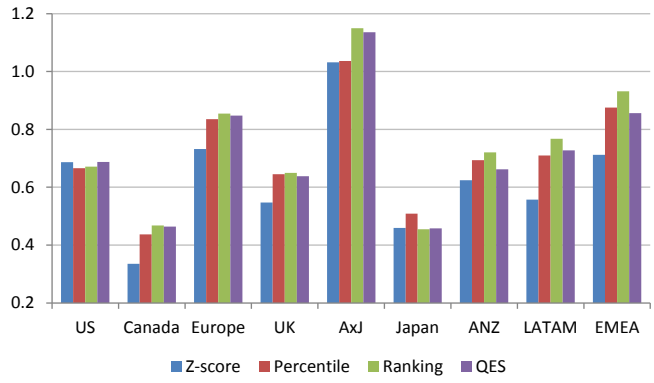
For each of the above factors, we apply three different transformation techniques, at each month end, within each of the nine regions:

- Z-score with winsorization (at ± 3 standard deviation)
- Percentile
- Ranking
- Our proprietary transformation

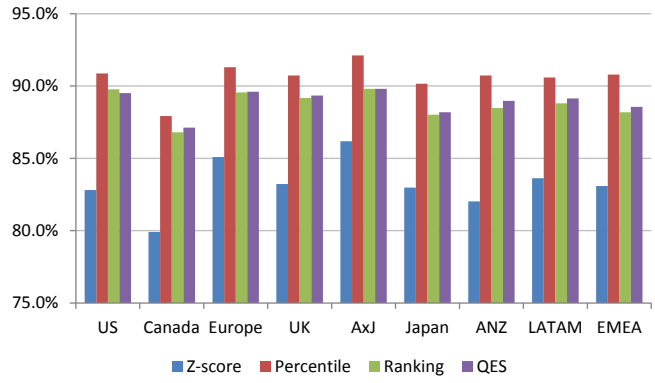
As shown in Figure 51 (A), across almost all nine regions, the three robust normalization approaches (percentile, ranking, and QES) outperform the traditional z-score model. Furthermore, the three procedures also produce significantly more stable models with higher signal autocorrelation (i.e., model predictions have smaller changes from month to month), which yields lower portfolio turnover.

Figure 51 The Impact of Data Transformation on Model Performance

A) Performance (Risk-Adjusted IC)



B) Model Turnover (Signal Autocorrelation)



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

MISSING VALUES

It is unrealistic to expect all factors to have 100% coverage for all companies in our universe, especially when we deal with small cap stocks, international firms, and/or unconventional datasets. How to deal with missing values is relevant and important to quantitative modeling.

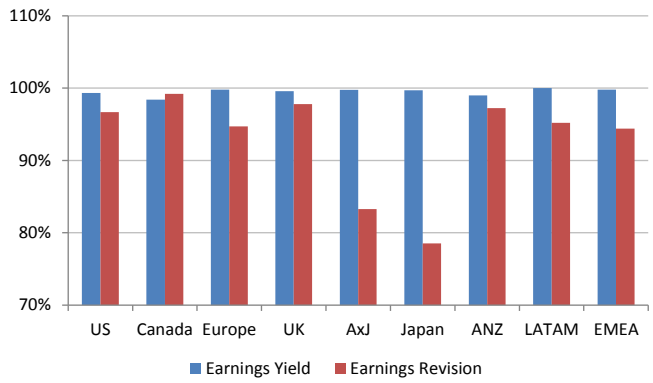
Factors that are based on financial statements and market data are likely to have good coverage, while signals that are derived from sell-side analysts and other unique information may have many missing values. As shown in Figure 52 (A), for example, earnings yield data has almost 100% coverage in all regions, while earnings revision suffers from almost 20% missing values in Japan (see Figure 52 B) and AxJ.

There are a few alternative approaches to adjust missing data:

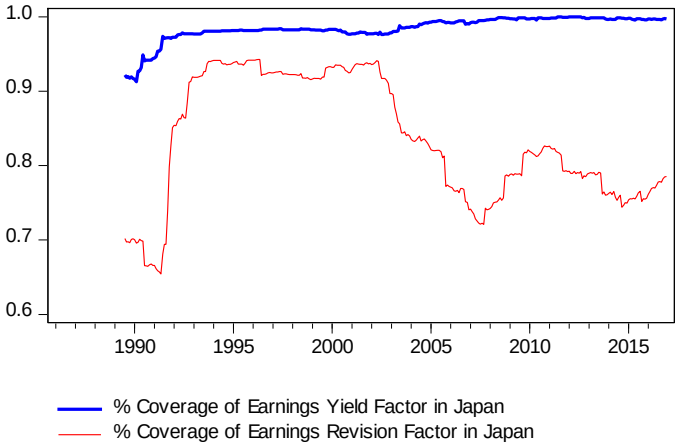
- Removing observations with missing value
- Filling missing values with sample mean/median
- Statistical multiple imputation
- Machine learning

Figure 52 The Coverage of Earnings Yield and Earnings Revision

A) Coverage by Region



B) Coverage in Japan



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

Removing Observations with Missing Values

The easiest way to deal with missing values is to ignore those stocks that we do not have data. For single factor backtesting, this approach is not only the most commonly used, but also the one produces the least biased result. However, as we start to combine factors into multi-factor models, this approach creates a serious problem. If we remove all observations with missing values from any of the underlying factors, we may end up with substantially smaller sample, which leads to information loss and reduction in breadth.

It is also common among investors to ignore factors with extraordinarily high percentages of missing values, but we run the risk of removing highly effective signals that work particularly well for a small universe of stocks.

Filing Missing Values with Sample Mean/Median

The most traditional approach is to fill missing values with sample mean (or median). If we normalize our data using a z-score type of transformation, the mean/median of our sample is close to zero. Therefore, we can simply assign a number of zero to all missing values. This approach is reasonable for most linear models, as zero or mean imputed values have no (or very limited) impact on the estimated coefficients and the subsequent predictions.

A more refined approach is first to divide our universe into buckets along risk dimensions (e.g., by country, sector) and then to replace missing scores with the mean/median of those corresponding bucket.

Statistical Multiple Imputation

Missing value imputation is a well-documented field in statistics. The most common model is called Predictive Mean Matching (PMM)[33]. In a high level, let's say we have $x_1, x_2, \cdots, x_K$ variables. To impute the missing values in x_1, we regress x_1 on the other variables x_2 to x_K. Similarly, we fill the missing values in x_2, we regress it on other variables $x_1, x_3, \cdots, x_K$. The regression can be simple OLS, logistic (for binary variable), Bayesian polytomous regression for categorical variables, etc.

In practice, multiple imputed data sets are generated. These data sets differ only in the imputed missing values. Then we proceed to our typical modeling process. Finally, we aggregate our model coefficients or predictions over the multiple samples.

When we apply the multiple imputation technique in factor data, we attempt to proxy the missing data using a linear combination of the other factors. Please note that we are not suggesting that the factor of interest is a linear combination of other factors – in that case, we have a multi-collinearity problem. Rather, it implies that filling the missing values with a linear function of other factors will result in minimum distortion to our model estimation.

Machine Learning Algorithm to the Rescue?

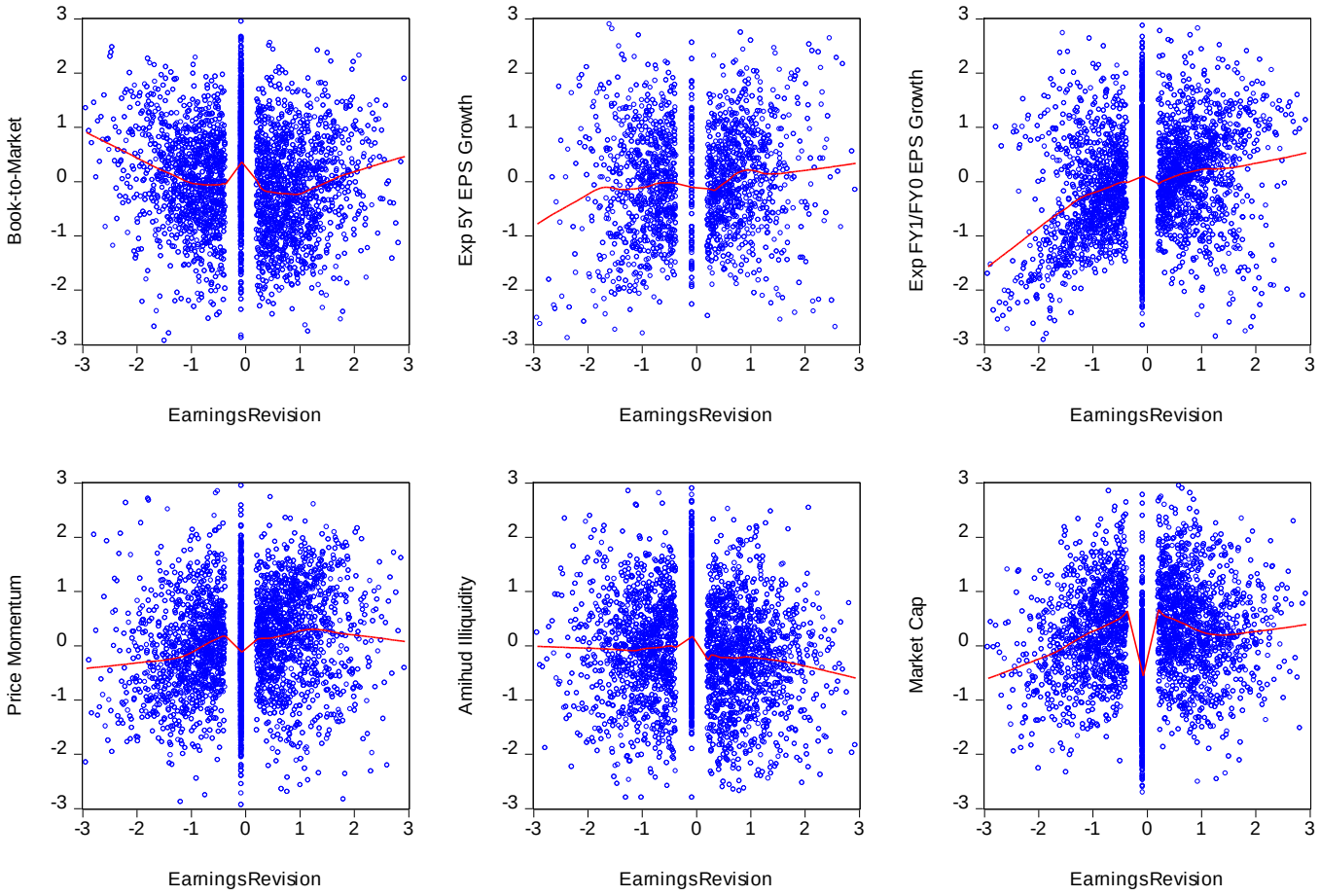
Machine learning algorithms can identify complex relationships better than traditional regression techniques. The simplest approach is the K-Nearest Neighbor (KNN) model, introduced by Troyanskaya et al [2001].

There are also more sophisticated techniques. For example, a random forest model (see Breiman [2001]) can be fitted for each variable. Then the random forest model can predict missing values in the variable with the help of the observed values from other variables. One of the major benefits of using the random forest algorithm is that it produces an OOB (Out-of-Bag) imputation error estimate for each variable; therefore, we can assess the goodness-of-fit for each factor.

As shown in Figure 53, the relationship between earnings revision and other common factors in AxJ appears to be loose and non-linear. Therefore, a nonlinear random forest algorithm might be able to capture the pattern in missing values better than a traditional linear multiple imputation model.

[33] Details about PMM can be found in Schafer [1997].

Figure 53 The Relationship between Earnings Revision and other Common Factors in AxJ



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

An Example

To study the impact of missing value imputation and compare various techniques, let's use a simple three-factor model (equally weighting the three factors) for AxJ:

- Value (trailing earnings yield)
- Price Momentum (12M total return)
- Analyst Sentiment (3M EPS revision)

We compare the performance of the model with the following missing value imputation methods for the above three factors. The coverage of value and momentum factors is almost 100%; therefore, the missing value imputation is mostly applied to analyst sentiment data (>15% missing values on average). We start the backtesting from 1997 and refit each of the missing value handling models every month.

- Removing missing values. We first remove all stocks with missing values in any of the three factors. The final multi-factor model is constructed on only stocks without any missing values.

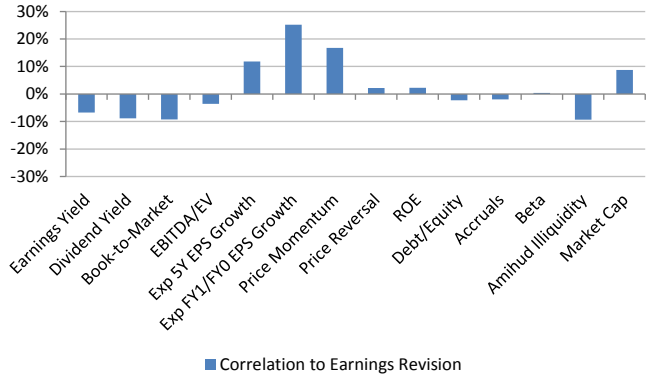
- Filling with Sample Mean. We first normalize all three factors AxJ, using our proprietary normalization algorithm discussed in the previous section. Then missing values are replaced with sample mean (i.e., zero), at each month end.
- Filling with Sector Mean. This is similar to the second approach above, but rather than filling missing data with sample mean, we use sector mean.
- PMM (Predictive Mean Matching). For each region, at each month end, we use a 12-month rolling window by stacking all data together. Then we perform the PMM imputation to fill missing values. The PMM imputation is based on 15 common factors and all factors are transformed to standard normal distribution first. We generate five imputed data sets. The final alpha is the simple average of the five data sets.
- Random Forest. For each region, at each month end, we fit a random forest model for each of the three factors, using 14 other common factors and all factors are transformed to standard normal distribution first. Then we use the fitted random forest model to "predict" the missing values, which are used in the final three-factor model.

Most factors are only modestly correlated to earnings revision in AxJ (see Figure 54 A). The most correlated factor is Consensus FY1/FY0 EPS growth, at around 25%. This is intuitive, because both factors are derived from sell-side analyst consensus. However, when earnings revision data is missing, most likely the earnings growth data is also missing. Therefore, the PMM approach depends on other correlated factors (e.g. price momentum and size) to impute the missing values for earnings revision.

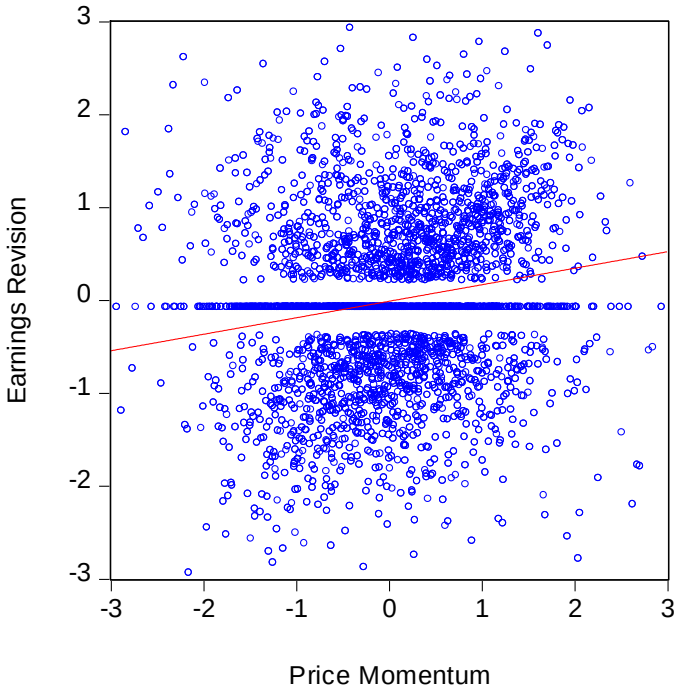
Figure 54 (B) plots the relationship between (non-missing) earnings revision and price momentum in AxJ. The statistical fit is very modest at best. Figure 54 (C) shows one set of imputed earnings revision using a linear PMM imputation algorithm. Clearly, it preserves the linear relationship between earnings revision and price momentum very well, albeit it also uses all other factors. Lastly, the random forest algorithm seems to capture some nonlinear pattern between earnings revision and price momentum.

Figure 54 The Comparison of Missing Value Imputation Techniques

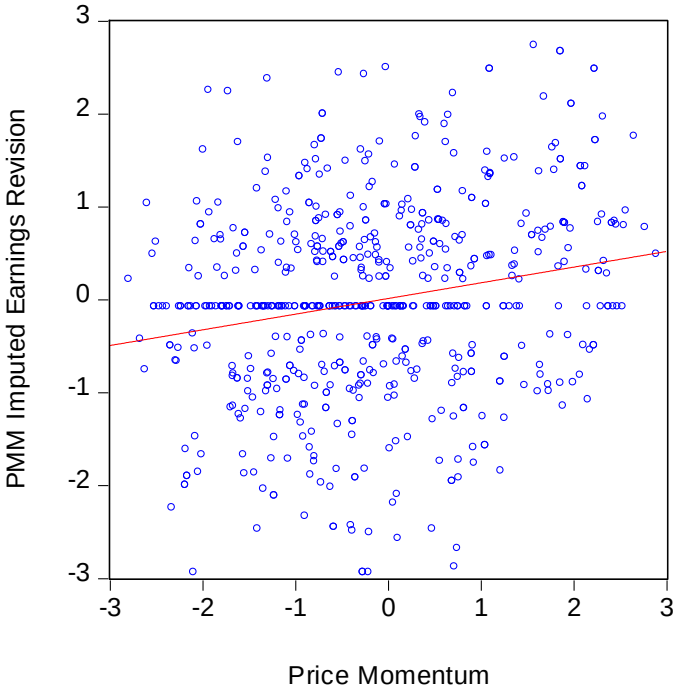
A) Correlation with Earnings Revision



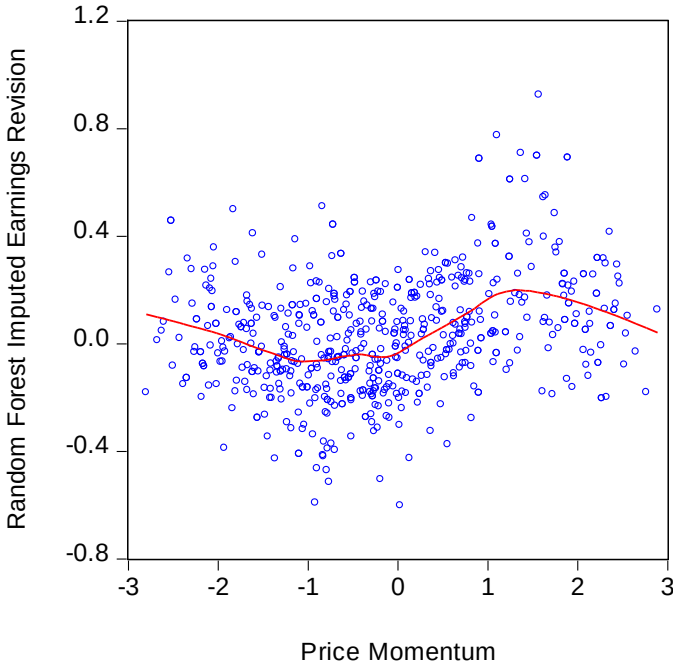
B) Price Momentum versus Earnings Revision



C) Price Momentum vs Imputed Earnings Revision, PMM



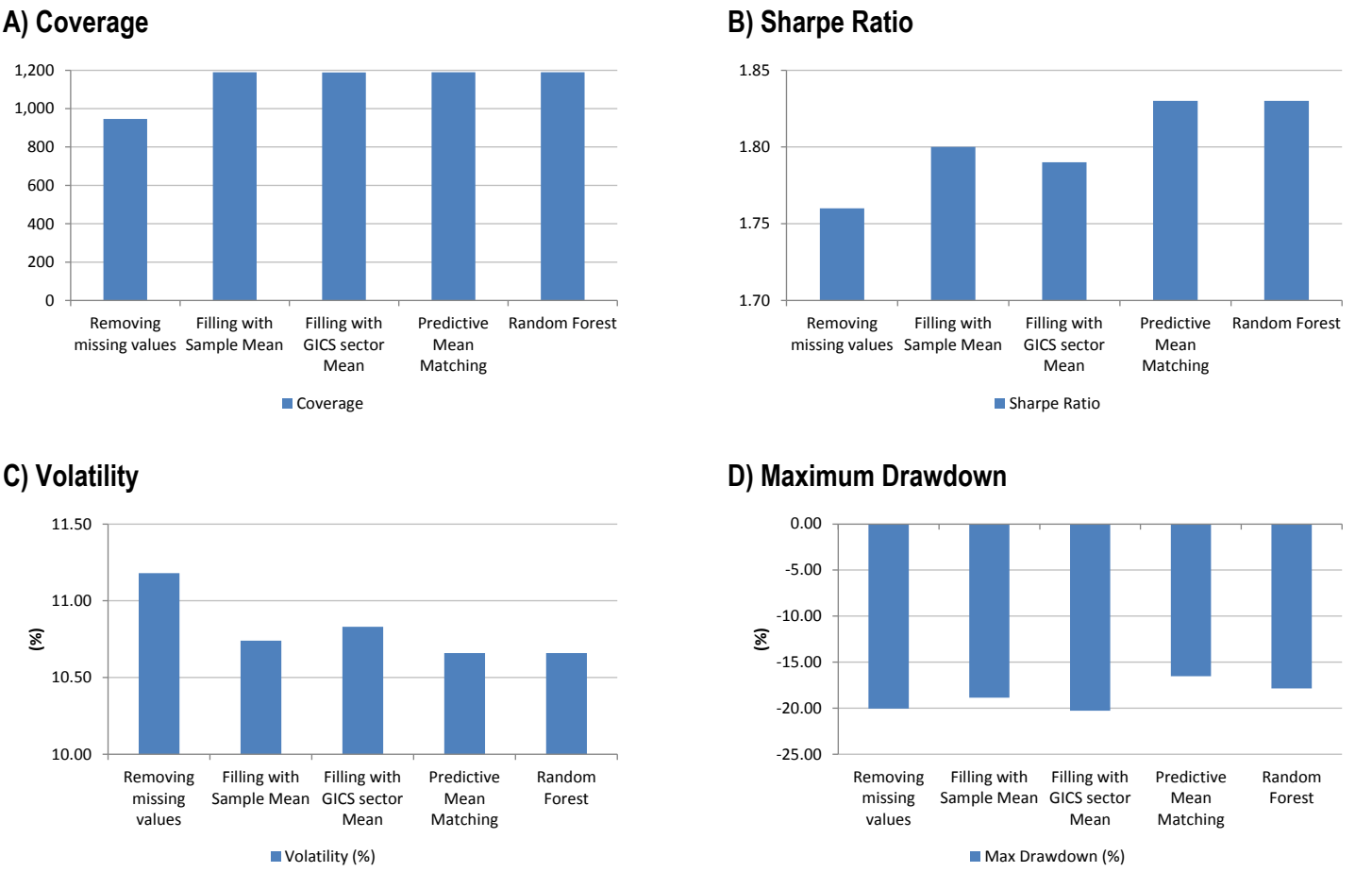
B) Price Momentum vs Imputed Earnings Revision, Random Forest



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

For each of the above five missing value handling approaches, we conduct a monthly backtesting, from 1997 until present. The Sharpe ratio, volatility, and maximum drawdown are based on a long/short quintile portfolio of the three-factor model in AxJ. The two statistical missing value imputation methods – PMM and random forest deliver similar performance. More importantly, these two sophisticated techniques increase our coverage, boost Sharpe ratio, and reduce volatility and downside risk, compared to more naïve approaches (see Figure 55).

Figure 55 Comparing Missing Value Handling Techniques



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

INTERNATIONAL MATTERS

As we move on from domestic equity to the international markets, we face additional complexities. We need to decide on how to treat currency in our modeling and whether to hedge or not to hedge our currency exposures. We have to choose between building a separate model for each country and combining multiple countries into a regional model.

Currency

When we compare companies from different countries, we need to take into account of currency adjustments. There are two currencies that we need to worry about: the reporting currency and the trading currency.

A company may have different currencies for financial reporting and trading. For example, BHP Billiton, an Anglo-Australian mining company, is headquartered in Australia, dual listed in both Australia's ASE and London (LSE). The Australia entity is part of the main Australian ASE 200 index and traded in Australian dollar (AUD), while the UK stock is part of the flagship FTSE 100 index, denominated in British Pound (GBP). Interestingly, the company's annual reports are presented in USD.

When we conduct currency translations, we need to apply different adjustments for financial statements from market data:

- Flow items, such as income statement and statement of cash flow items, need to be adjusted using the average exchange rate over the reporting period;
- Period end (stock) items, such as balance sheet items, need to be translated using the period-end exchange rate;
- Market data are generally adjusted using the spot exchange rate on the pricing date

We could compute a factor for all companies using the same currency, e.g., USD. Alternatively, factors can be calculated using each company's own local currency. When we compute factors that combine financial statement data and market data, we need to be particularly careful. For example, to compute earnings yield, we divide share price by earnings per share (EPS). If a company uses one currency for financial reporting (EPS) and another one for trading (price), the earnings yield number is meaningless if they are denominated in difference currencies.

When we conduct our backtesting, we also need to choose the currency for our strategy return calculation. Normally, the assumption is that, if a manager fully engages in full currency hedge, it is appropriate to use local currency to compute returns. On the other hand, if a manager does not hedge currency at all, she may want to use USD (or the main currency of her portfolio).

When we conduct backtesting, we have four combinations of currency treatment:

- Factor and stock both computed in local currency
- Factor computed in USD (or other common currency) and stock return denominated in local currency
- Factor computed in local currency and stock return denominated in USD (or other common currency)
- Factor and stock return both computed in USD (or other common currency)

To assess the impact of currency on backtesting, we perform a backtesting with three factors:

- Earnings yield (EPS/Price), where EPS are from financial statements and translated at the average exchange rate over the fiscal period; and price is market data and translated at the spot exchange rate

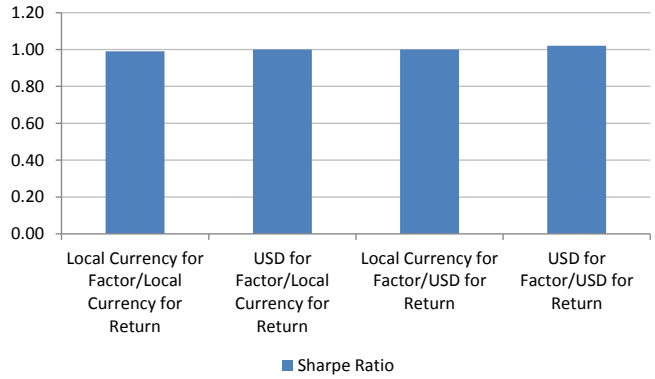
- ROE (Net Income/Book Value of Equity), where both denominator and numerator are from the financial statements; net income is a flow item and translated using the average exchange rate over the fiscal period; and book value of equity is a stock item and translated using the spot exchange rate at the end of the fiscal period
- Price Momentum, computed as the total return over the past 12 months, excluding the most recent one month, using spot exchange rates at the beginning and end periods

Our investment universe is all stocks in all the countries in developed Europe excluding UK. We perform a standard monthly rebalanced backtesting, from 1991 until present.

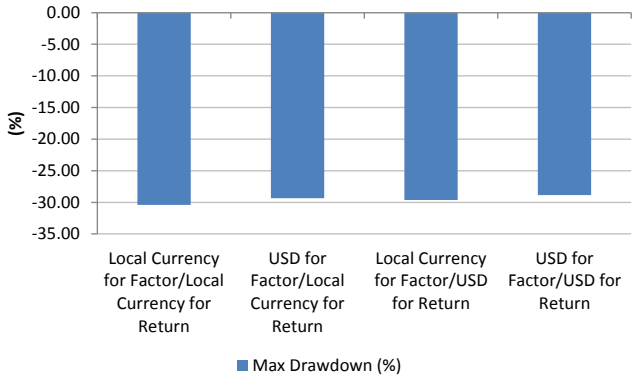
Interestingly, across the three factors, the four sets of currency adjustments produce almost identical results (see Figure 56). For our backtesting, unless it is separately disclosed, we normally use local currency to compute both factors and returns.

Figure 56 The Impact of Currency on Backtesting

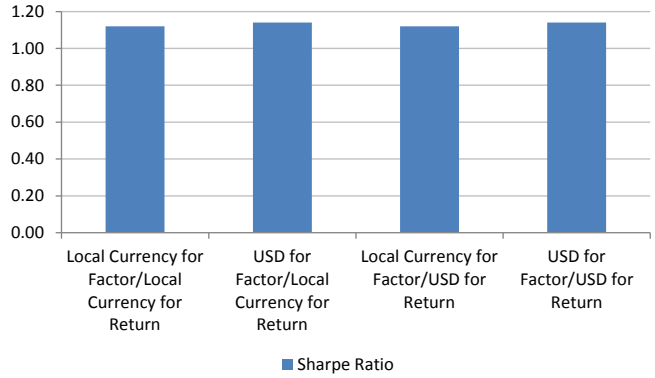
A) Earnings Yield, Sharpe Ratio



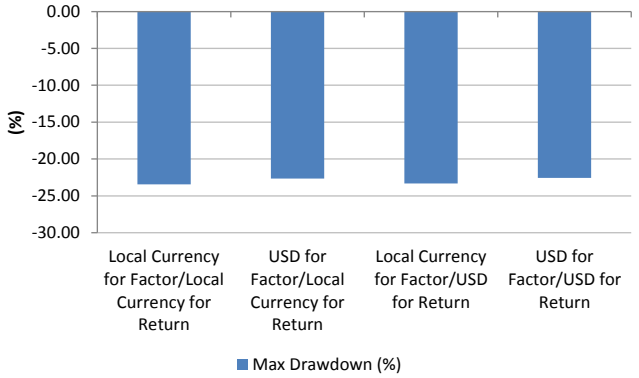
B) Earnings Yield, Maximum Drawdown



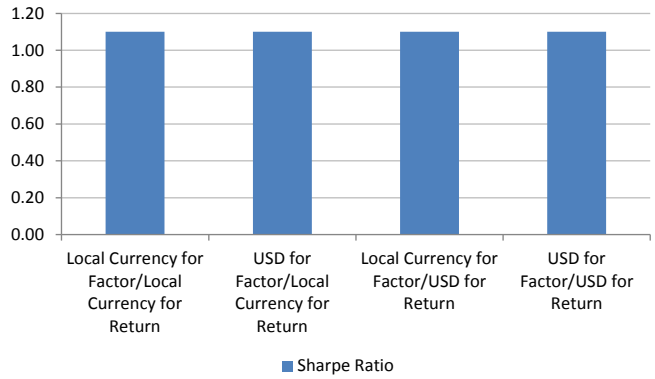
C) ROE, Sharpe Ratio



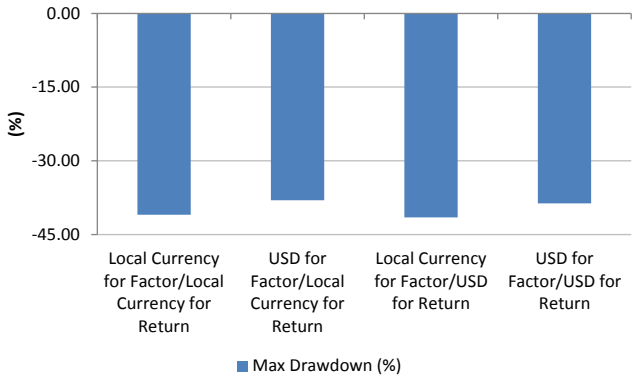
D) ROE, Maximum Drawdown



E) Price Momentum, Sharpe Ratio



F) Price Momentum, Maximum Drawdown



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

Can We Compare Companies from Different Countries?

Although most countries use IFRS, there are a few major countries with distinctive local GAAP. The most obvious example is of course, the US. Despite the efforts made by the FASB and IFRS to reconcile the two sets of accounting principles, there remain significant differences in a number of important areas such as fair value accounting and even revenue recognition. Other major countries such as, China, India, and Japan still use their own country GAAP, but the local GAAP is increasingly converging to the IFRS.

Some sectors are more global, such as energy and materials, as the underlying commodities prices are driven by global supply and demand. Some sectors tend to be more local, such as utilities. These sectors can be highly regulated domestically, operated mostly within a country, and primarily service local customers.

Some countries are more closely linked together. The classic example is the European Union, including most developed European countries even those are not officially in the EU. The UK, however, has always been only half integrated in the EU. The recent Brexit vote added even more uncertainty.

The decision of country-specific or regional (global) model is a decision of sample size versus relevance. In a perfect world, the backtesting universe should include stocks that are as similar as possible, e.g., in the same country, sector, size, and risk bucket[34]. However, in practice, splitting the universe too many ways will result in a tiny sample.

For demonstration purpose, we choose three factors in five countries, for a standard monthly backtesting. The three factors are:

- Earnings Yield
- ROE
- Price Momentum

We choose the following five countries:

- France
- Germany
- UK
- Australia
- New Zealand

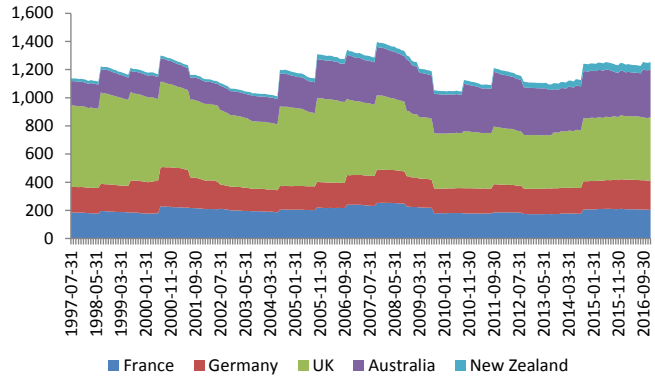
We want to understand whether factors are more correlated within the same country, or the same factor is more likely to move in unison across countries. Furthermore, are factors more correlated between Germany and France than with UK, and maybe even less so compared to Australia and New Zealand? We also need to take size (i.e., number of stocks for each country) into account. As shown in Figure 57 (A), there are far more stocks in the UK than the other countries. Similarly, New

[34] This is the philosophy of "contextual analysis" by Qian, Hua and Sorensen [2007]. However, since they study the US equity market, there are enough stocks in both value and growth sub-universes.

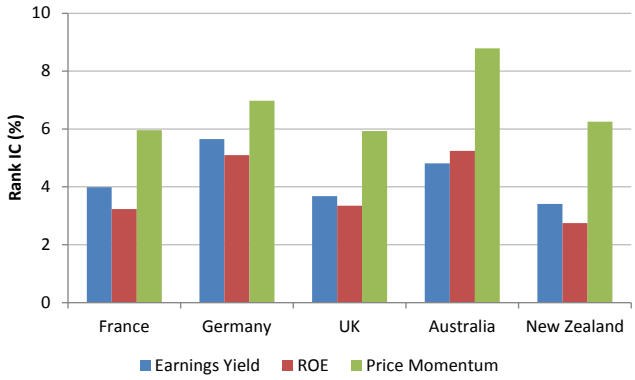
Zealand is substantially smaller than the other four markets. Figure 57 (B) demonstrates the performance of the three factors in five countries.

Figure 57 Three Common Factors in Five Countries

A) Coverage by Country



B) Average Rank IC, by Country



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

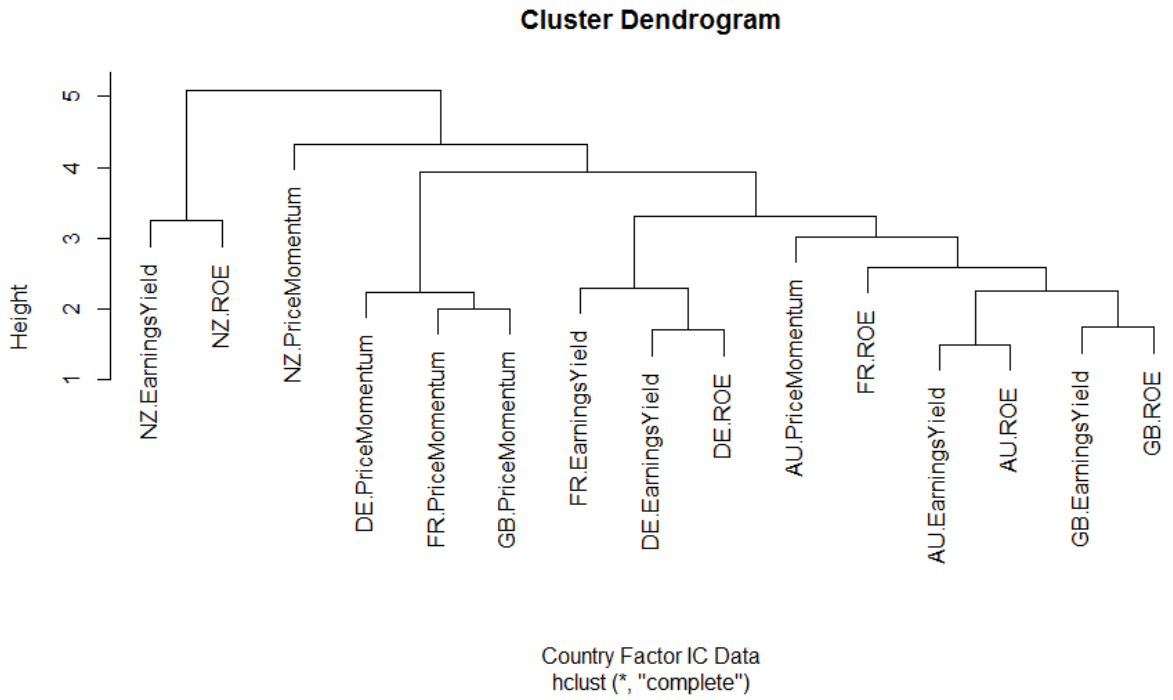
We perform a cluster analysis on the three factors in five countries. As shown in Figure 58, we detect a few interesting patterns:

- Factors are highly correlated between France and Germany;
- Factors in the UK are loosely associated with the ones in France/Germany, but not as strong as factors between France and Germany;
- Factors in Australia and New Zealand behave differently from signals in Europe

Our empirical backtesting supports our breakdown of the world, as we:

- Combine all developed countries in Europe in one region
- Separate UK from continental Europe
- Merge Australia and New Zealand in one universe

Figure 58 A Cluster Analysis on Factor Performance (Rank IC) across Countries



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

FORTHCOMING RESEARCH

Due to size limit, this paper only covers the first part of systematic investing. In the next few weeks, we will publish the other three key components:

- *Part II: Signal Research and Multifactor Models* – From Data to Knowledge
- *Part III: Machine Learning, Style Rotation, and the Next Frontier in Systematic Investing*
- *Part IV: Risk, Portfolio Construction, Trade Execution, and Performance Attribution* – From Theory to Practice

In addition to the four-part introduction of Big Data and Machine learning in global equity investing, we are also working on a number of issues:

- From Nowcasting to Forecasting – Economics and Portfolio Strategy in the New Age
- Industry-Specific Models in Global Banking and Insurance Industries
- Accounting Quality, Fraud Detection, and Corporate Governance
- Factors based on Alternative Data Sources
- Machine Learning in Global Stock Selection

BIBLIOGRAPHY

Box, G.E.P., and Cox, D.R. [1964]. "An Analysis of Transformations", *Journal of the Royal Statistical Society*, series B 26(2): 211-252.

Breiman, L. [2001]. "Random Forests", *Machine Learning*, 45, 5-32.

Cahan, R., and Luo, Y. [2013]. "Standing Out from the Crowd: Measuring Crowding in Quantitative Strategies", *Journal of Portfolio Management*, Summer 2013.

Fama, E., and French, K.R. [1993]. "Common Risk Factors in the Returns on Stocks and Bonds", *Journal of Financial Economics* 33(1): 3-56.

Fama, E., and French, K.R. [1996]. "Multifactor Explanations of Asset Pricing Anomalies", *Journal of Finance*, Vol. 51, 55-84.

Grinblatt, M., Masulis, R., and Titman, S. [1984]. "The Valuation Effects of Stock Splits and Stock Dividends", *Journal of Financial Economics*, 13, 461-90.

Jegadeesh, N., and Titman, S. [1993]. "Returns to buying winners and selling losers: implications for stock market efficiency", *Journal of Finance*, Vol. 28, No. 1, 63, March 1993.

Khandani, A.E., and Lo, A.W. [2007]. "What Happened to the Quants in August 2007", MIT Working Paper.

Mohanram, P.S. [2005]. "Separating Winners from Losers among Low Book-to-Market Stocks Using Financial Statement Analysis", *Review of Accounting Studies*, 10, 133-170.

Piotroski, J.D. [2002]. "Value Investing: The Use of Historical Financial Statement Information to Separate Winners from Losers", University of Chicago GSB Selected Paper 84.

Schafer, J.L. [1997]. *Analysis of Incomplete Multivariate Data*, Chapman & Hall/CRC, London.

Sloan, R.G. [1996]. "Do Stock Prices Fully Reflect Information in Accruals and Cash Flows about Future Earnings", *The Accounting Review* 71(3), 289-315.

Stone, M. and Brooks, R. [1990]. "Continuum Regression: Cross-Validated Sequentially Constructed Prediction Embracing Ordinary Least Squares, Partial Least Squares, and Principal Component Regression", *Journal of the Royal Statistical Society*, Series B, 52, 237-269.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P. Hastie, T., Tibshirani, R., Botstein, D., Altman, R. [2001]. "Missing Value Estimation Methods for DNA Microarrays", *Bioinformatics*, 17(6), 520-525.

DISCLOSURE SECTION

Analyst Certification:
The analyst of Wolfe Research primarily responsible for this research report whose name appears first on the front page of this research report hereby certifies that (i) the recommendations and opinions expressed in this research report accurately reflect the research analysts' personal views about the subject securities or issuers and (ii) no part of the research analysts' compensation was, is or will be directly or indirectly related to the specific recommendations or views contained in this report.

Other Disclosures:
Wolfe Research, LLC does not assign ratings of Buy, Hold or Sell to the stocks it covers. Outperform, Peer Perform and Underperform are not the respective equivalents of Buy, Hold and Sell but represent relative weightings as defined above. To satisfy regulatory requirements, Outperform has been designated to correspond with Buy, Peer Perform has been designated to correspond with Hold and Underperform has been designated to correspond with Sell.

Wolfe Research Securities and Wolfe Research, LLC have adopted the use of Wolfe Research as brand names. Wolfe Research Securities, a member of FINRA (www.finra.org) is the broker-dealer affiliate of Wolfe Research, LLC and is responsible for the contents of this material. Any analysts publishing these reports are dually employed by Wolfe Research, LLC and Wolfe Research Securities.

The content of this report is to be used solely for informational purposes and should not be regarded as an offer, or a solicitation of an offer, to buy or sell a security, financial instrument or service discussed herein. Opinions in this communication constitute the current judgment of the author as of the date and time of this report and are subject to change without notice. Information herein is believed to be reliable but Wolfe Research and its affiliates, including but not limited to Wolfe Research Securities, makes no representation that it is complete or accurate. The information provided in this communication is not designed to replace a recipient's own decision-making processes for assessing a proposed transaction or investment involving a financial instrument discussed herein. Recipients are encouraged to seek financial advice from their financial advisor regarding the appropriateness of investing in a security or financial instrument referred to in this report and should understand that statements regarding the future performance of the financial instruments or the securities referenced herein may not be realized. Past performance is not indicative of future results. This report is not intended for distribution to, or use by, any person or entity in any location where such distribution or use would be contrary to applicable law, or which would subject Wolfe Research, LLC or any affiliate to any registration requirement within such location. For additional important disclosures, please see www.wolferesearch.com\disclosures.

The views expressed in Wolfe Research, LLC research reports with regards to sectors and/or specific companies may from time to time be inconsistent with the views implied by inclusion of those sectors and companies in other Wolfe Research, LLC analysts' research reports and modeling screens. Wolfe Research communicates with clients across a variety of mediums of the clients' choosing including emails, voice blasts and electronic publication to our proprietary website.