Quantitative Methods

Backtesting and Simulation

by Yin Luo, CPA, PStat, CFA, and Sheng Wang

Yin Luo, CPA, PStat, CFA, is at Wolfe Research LLC (USA). Sheng Wang is at Wolfe Research LLC (USA).

LEARNING OUTCOMES

Mastery	*The candidate should be able to:*
☐	**a.** describe objectives in backtesting an investment strategy;
☐	**b.** describe and contrast steps and procedures in backtesting an investment strategy;
☐	**c.** interpret metrics and visuals reported in a backtest of an investment strategy;
☐	**d.** identify problems in a backtest of an investment strategy;
☐	**e.** describe different ways to construct multifactor models;
☐	**f.** compare methods of modeling randomness;
☐	**g.** evaluate and interpret a scenario analysis;
☐	**h.** contrast Monte Carlo and historical simulation;
☐	**i.** explain inputs and decisions in simulation and interpret a simulation; and
☐	**j.** demonstrate the use of sensitivity analysis.

INTRODUCTION 1

This reading provides an overview of backtesting and simulation of investment strategies. Backtesting and related techniques enable investment practitioners to simulate the performance of investment strategies (especially quantitative strategies) using historical data or data derived from the distributions of historical data, to generate test results, and to analyze risk and return, without investing any real capital in the strategies.

The rise of big data and the increase in computing power have spurred the development and spread of quantitative investing. Almost every major data vendor has available tools that make systematic backtesting and simulation increasingly accessible.

Off-the-shelf software allows backtesting and simulation of endless combinations of possible investment strategies, formulation of multifactor models, and construction of investable portfolios. Developing quantitative investment strategies may appear relatively straightforward, but in reality, it is not. However, understanding the steps and procedures, the implicit assumptions, the pitfalls, and the interpretation of results in backtesting and simulation is a prerequisite for proper utilization of these tools and successful development and implementation of investment strategies.

In a CFA Institute survey of nearly 250 analysts, portfolio managers, and private wealth managers on quantitative investment techniques, 50% of respondents reported that they had conducted backtesting of an investment strategy within the past 12 months of the survey date. This result underscores the importance of backtesting (and other simulation techniques) for investors in practice, and this reading is a starting point on the journey to building this core professional competency.

2 THE OBJECTIVES OF BACKTESTING

a Describe Objectives in Backtesting an Investment Strategy

Backtesting is the process that approximates the real-life investment process, using historical data, to assess whether the strategy would have produced desirable results. Backtesting offers investors some comfort as to whether their investment strategies and analytical models would have performed well historically. More importantly, it allows investors to refine and optimize their investment process.

Backtesting has been widely used in the investment community for many years. Although backtesting fits quantitative and systematic investment styles more naturally, it has also been heavily used by fundamental managers. As long as you use any form of stock screening, you may want to first backtest and see whether your selection criteria and process indeed add any incremental excess return.

Notably, not all strategies that have performed well in a backtest will continue to produce excess returns in live investing. Although in theory a model that does not show much predictive power in backtesting could still deliver excess return in the future, such a model is unlikely to be sufficiently convincing for portfolio managers to implement.

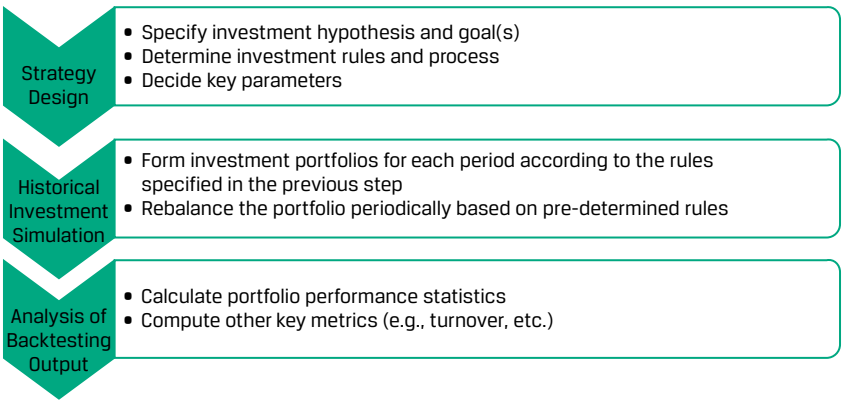
Importantly, the implicit assumption in backtesting is that the future will as least somewhat resemble history. Therefore, if we implement our investment strategies that have performed well in backtesting, we may expect similar performance going forward. The reality, however, as we will show, is far more complicated. There are a number of factors—some are under the investment managers' control and some are not—that may completely disrupt our investment process.

3 THE BACKTESTING PROCESS

b Describe and Contrast Steps and Procedures in Backtesting an Investment Strategy

Backtesting typically follows the steps and procedures of strategy design, historical investment simulation, and analysis of backtesting output. These steps and associated procedures are set out in Exhibit 1. We will explain each of these steps and procedures in detail in the following sections.

Exhibit 1 Backtesting Flowchart



Source: Wolfe Research Luo's QES.

3.1 Strategy Design

The first step in backtesting is to specify the investment hypothesis and goals. For active strategies, the goal is typically to achieve superior risk-adjusted excess returns. However, in many situations, downside risk management is also critical. Furthermore, depending on how the strategies will be deployed—quantitatively or in a fundamental investment process—other considerations, such as portfolio turnover, portfolio concentration, and investment horizon, should also be taken into account.

Common equity investment strategies often use factor-based models. A factor can be broadly considered as any variable that is believed to be useful in ranking stocks (in terms of attractiveness) for investment and in predicting future returns or risks. Value and momentum (which we will discuss later) are two classic examples of factors; factors represent unique sources of risk with sound economic foundations. A factor-based strategy aims to identify significant signals that drive stock prices, with the aim of forming investable portfolios to outperform the market.

As mentioned earlier, although fundamental portfolio managers do not invest solely based on factors, they typically perform some type of stock screening to identify potential investment opportunities. This is especially the case for small-cap and global equity managers, because their investment universes are too large to conduct detailed fundamental analysis on all companies. Most commonly known investment styles can be proxied by factors. For example, the return of most value funds is highly correlated to the performance of valuation factors, such as book-to-market ratio, earnings yield (earnings-to-price ratio), and price-to-sales ratio. Therefore, backtesting such factors is a critical element for managers to improve and refine their investment process.

In this section, we will use a basic value signal and a classic price momentum factor to demonstrate how backtesting can be conducted.

3.1.1 *Value*

Value investing is based on the concept of buying undervalued stocks whose prices would eventually rise and converge toward their intrinsic values. Academic literature has a long history of documenting the value phenomenon. Early researchers found that stocks with low price-to-earnings ratios (P/Es) or high earnings yield (i.e., the inverse of P/E) tend to provide higher returns. Fama and French (1993) formally outlined value investing by proposing the use of the book-to-market ratio as a way to differentiate value stocks from growth stocks.

In practice, value factors can be based on almost any fundamental performance metrics of a company, such as dividends, earnings, cash flow, EBIT, EBITDA, and sales. Most valuation ratios can also be computed using either historical (called trailing) or forward-looking (i.e., expected future) metrics. In this section, we use a simple metric as our factor—trailing earnings yield:

$$\text{Trailing earnings yield} = \frac{\text{Trailing 12-month EPS}}{\text{Current share price}}. \quad (1)$$

Following industry practice, we expect "cheap" stocks with more attractive valuations (i.e., with higher earnings yields) to outperform "expensive" stocks (i.e., with lower or even negative earnings yields).

3.1.2 *Momentum*

Researchers have found a strong price momentum effect in almost all asset classes. Jegadeesh and Titman (1993), in a study of the US market, first documented that stocks that are the "winners" over the last 12 months tend to outperform past losers over the next 2–12 months. The authors also found there is a short-term reversal effect—that is, stocks with a high price momentum over the last month tend to underperform in the next 2–12 months. As a result, practitioners often define the price momentum factor using the following equation, which focuses on the total return in the past 12 months, excluding the most recent month:

$$\text{Price momentum} = r_{-12:-1}. \quad (2)$$

Note that the reason to exclude the most recent month's return is to account for the short-term mean-reversal effect.

3.2 Rolling Window Backtesting

Once a potential investment strategy is designed, the investment manager needs to perform a series of backtests to assess the factor's performance and effectiveness. Some investors argue that the most important test is the "intuition" test: Does the factor make intuitive sense? Is there an underlying economic rationale? A factor can often pass statistical backtesting, but if it does not make sense and a reasonable justification for the factor's efficacy is lacking, then the manager may have fallen into the data-mining trap. The economic intuition test, however, lacks theoretical rigor and can be easily abused. It is often difficult (and subjective) to differentiate what counts as being intuitive and what does not.

Investors, therefore, always need to remind themselves that what appears to be impressive performance from backtesting does not guarantee the same returns in the future.

In a typical backtest, researchers first form their investment hypothesis—for example, stocks with higher earnings yield should earn higher subsequent returns—and then determine their investment rules and processes. Next, they collect the necessary data—in this case, historic earnings yield and return for each stock. In the simplest type of backtest, researchers divide the historical data into two subsamples. The first few years of data, the first subset, is used to train the model (i.e., in-sample/training data). The trained model is then tested using the second subset of data (i.e., out-of-sample/testing data).

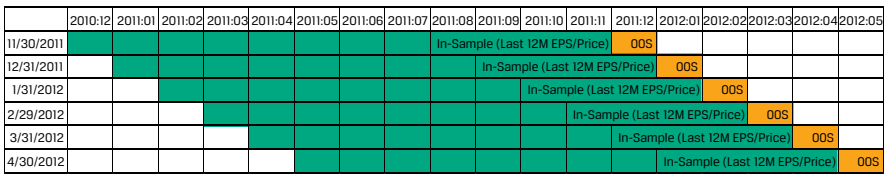
In the **rolling window backtesting** methodology, instead of dividing the data into only two samples, researchers typically use a rolling window (or walk-forward) framework, fit/calibrate factors or trade signals based on the rolling window, rebalance the portfolio periodically (i.e., after each period), and then track the performance over time. In this case, backtesting is a proxy for actual investing. As new information arrives, investment managers re-adjust their models and rebalance their stock positions. To

avoid overfitting, in the backtesting, the key model methodology should be specified up front. Therefore, at each given point in time, the model is tuned according to its specification. Most backtesting procedures assume monthly model optimization and position rebalancing—that is, researchers repeat the same in-sample training/out-of-sample testing process on the last day of each month. Other common frequencies, such as daily, weekly, and quarterly, are also used in practice by investors. Note that these overlapping periods may not be considered fully independent statistically, and the rolling window methodology may be slow to pick up regime changes.

Exhibit 2 shows a snapshot of such a rolling window backtesting of the value (trailing 12-month earnings yield) factor. Assume that we start our backtesting on 30 November 2011. First, we compute every stock's trailing 12-month earnings yield, using EPS reported in the previous 12 months (i.e., from December 2010 to November 2011), divided by current stock price as of 30 November 2011. We can then form our investment strategy—for example, buying the top 20% of stocks with the highest earnings yield and shorting the bottom quintile of companies (the 20% of stocks with the lowest earnings yield). The performance of such a strategy can be assessed using returns in the next month, December 2011, which is the out-of-sample (OOS) period. The same process is repeated on 31 December 2011, and so on. In our simple example, we backtest our value strategy from November 2011 to April 2012. We then compute the average monthly return, volatility, Sharpe ratio, and drawdown of our strategy from the test results of the six OOS periods.

If the performance of the investment strategy in the out-of-sample periods is desirable and the strategy meets the economic intuition test, then it is generally considered successful. Otherwise, the strategy is rejected.

Exhibit 2 An Example of Rolling Window Backtesting of the Earnings Yield Factor



Source: Wolfe Research Luo's QES.

3.3 Key Parameters in Backtesting

Before we can perform a backtesting exercise, there are a few key parameters that must be specified, including the investment universe, the specific definition of stock return, the frequency of portfolio rebalancing, and the start and end dates for the backtest.

3.3.1 *Investment Universe*

First, we need to decide on the investment universe—that is, the universe of stocks in which we can potentially invest. Investment strategy performance can be vastly different between different countries, different sectors, and different contexts (e.g., value versus growth). While academic research typically uses the union of Compustat/Worldscope and CRSP[1] as the research universe, practitioners often use a well-known

1 CRSP (the Center for Research in Security Prices) provides high-quality data and security returns. The CRSP data series of New York Stock Exchange–listed stocks begins on 31 December 1925.

broad market index, such as the S&P 500 Index, the Russell 3000 Index, or the MSCI World, to define their investment universe. In this reading, unless it is otherwise specified, we use the Russell 3000 index to represent the US market, the S&P/TSX Composite Index to represent the Canadian market, the MSCI China A Index to represent the Mainland China market, and the S&P Global Broad Market Index (BMI) for markets covering all other countries.

3.3.2 *Stock Return*

Since the goal of backtesting is to proxy the real investment process, we compute stock returns by taking into account not only capital appreciation but also dividend reinvestment.

As we extend our investment universe from a single country to a global context, multiple complexities arise, such as currency, trading, and regulatory considerations. For example, we need to decide in what currency the return should be computed. The two most frequent choices are either to translate all investment returns into one single currency, typically the home-country currency (e.g., US dollar or euro) or to denominate returns in local currencies. The choice of currency in backtesting often depends on whether portfolio managers prefer to hedge their currency exposures. Managers who do not hedge their foreign exchange risk often choose to backtest using single-currency-denominated returns.

3.3.3 *Rebalancing Frequency and Transaction Cost*

Academic studies typically use investment returns calculated quarterly or annually, while practitioners mostly use a monthly frequency for their rolling windows for model optimization and portfolio rebalancing. Note that daily or even higher-frequency rebalancing can incur high transaction costs, and the price data will likely be biased by bid–ask spreads, asynchronous trading across different parts of the world, and missing days due to holidays in different countries.

In most standard backtesting, transaction costs are typically not included, because researchers want to first understand the "pure" alpha from the strategy. More importantly, transaction costs critically depend on the portfolio construction process; therefore, they vary significantly from one portfolio manager to another. For example, the following issues should be accounted for when estimating transaction costs:

- whether the strategy is implemented in a systematic approach or used as a stock screening process;
- whether the portfolio is long only or long/short market neutral;
- the size of the portfolio (i.e., assets under management);
- whether currency is hedged; and
- how trades are executed—via electronic trading, program trading, or single-stock trading.

In sum, transaction costs are critical to whether a strategy is profitable in practice, because many market anomalies simply disappear once costs are included.

3.3.4 *Start and End Date*

Everything else being equal, investment managers typically prefer to backtest their investment strategies using as long a history as possible for their model inputs. Using a longer history for asset prices, returns, factors, and so on, provides a larger sample; therefore, the investment manager would have a greater statistical confidence in the backtesting results. Conversely, however, since financial data are likely to be non-stationary with structural breaks, backtested performance using a long data history may not be very relevant looking forward. Our recommendation is to backtest as

far back as possible. However, managers may choose to weight recent periods more heavily compared with distant history. Importantly, researchers need to be aware that historical periods may not be relevant for their investment thesis (e.g., past periods of high inflation may be less relevant for a strategy based on a low-inflation environment). Later, in the coverage of scenario analysis, we will discuss how to deal with structural changes and regime shifts.

3.4 Long/Short Hedged Portfolio Approach

The most traditional and widely used method for implementing factor-based portfolios is the hedged portfolio approach, pioneered and formulated by Fama and French (1993). In this approach, after having chosen the factor to be scrutinized and having ranked the investable stock universe by that factor, the analyst divides the universe into groups referred to as quantiles (typically into quintiles or deciles) to form hypothetical portfolios. Stocks are either equally weighted or market capitalization weighted within each quantile. A long/short hedged portfolio is then formed by going long the top quantile (i.e., the one with the best factor scores) and shorting the bottom quantile (i.e., the one with the worst factor scores). The rolling window backtesting framework is then implemented and the portfolio is rebalanced periodically—for example, monthly. The performance of the long/short hedged portfolio is then tracked over time (i.e., for each successive out-of-sample period).

The long/short hedged portfolio is dollar neutral but not necessarily beta neutral. For example, the long portfolio formed using the book-to-market factor typically contains stocks with higher beta than the short portfolio, because cheap stocks are often more volatile than expensive stocks. As a result, the return from such a portfolio is partially due to exposures to the market. Similarly, the portfolio may also have exposures to other common factors. However, as a reasonable and straightforward approximation of the performance of the factor-based strategy, the hedged portfolio approach serves the purpose. We care about not only the average return of the portfolio but also the risk profile (e.g., volatility and downside risk). Therefore, most common performance measurement metrics, such as the Sharpe ratio, the Sortino ratio, and maximum drawdown, can be used to assess the efficacy of our investment strategy.

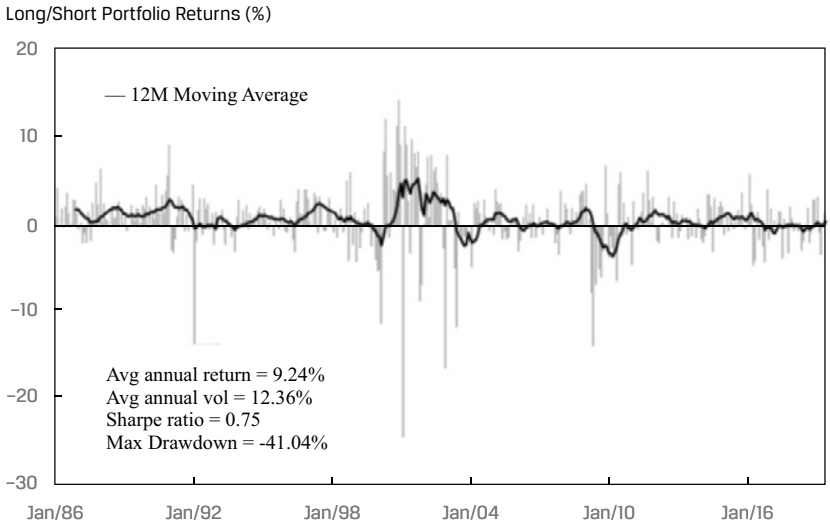
EXAMPLE 1

Performance of Earnings Yield Factor in the United States and Europe, 1986–2019

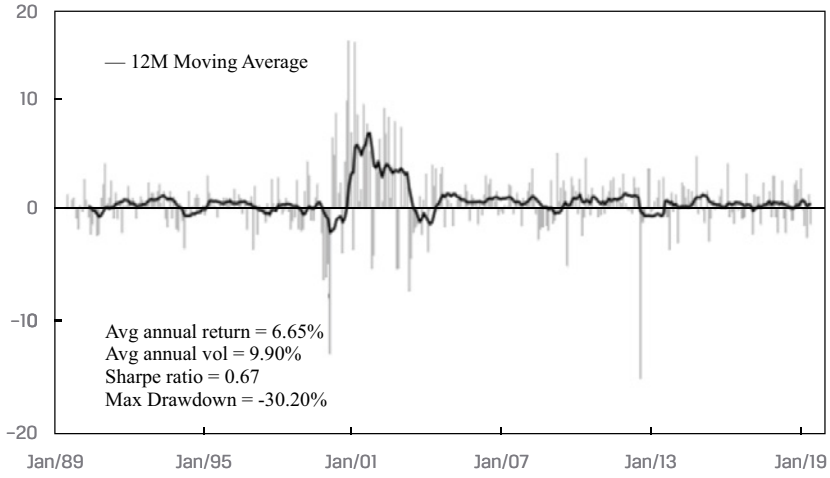
Panel A of Exhibit 3 shows the performance of the earnings yield factor in the United States from January 1986 to May 2019. The bars in the chart indicate the monthly portfolio returns generated from buying companies with the highest earnings yields, those in the top quintile, and shorting companies with the lowest/most negative earnings yields, those in the bottom quintile. Stocks in both long and short baskets are equally weighted. Panel B shows the performance of the earnings yield factor in Europe over the same time period.

Exhibit 3 Earnings Yield Factor, Long/Short Hedged Quintile Portfolio Returns (January 1986–May 2019)

A. US: Trailing Earnings Yield



B. Europe: Trailing Earnings Yield



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

Describe how the backtest performance of value investing, based on the earnings yield factor, in Europe compares with that in the United States over the period of 1986–2019.

Solution:

In the United States, the average annual return from the value investing strategy is about 9.2%, with a Sharpe ratio of 0.75, over the backtesting period (January 1986–May 2019). In Europe, the same investment strategy generated a significantly lower (by 250 bps) average annual return, about 6.7%, but with significantly lower volatility. Hence, the Sharpe ratio for the European strategy, 0.67, is close to that of the US strategy. In both markets, the maximum drawdown is just over three times the volatility of the strategy.

Therefore, as a long-term value strategy, the earnings yield factor offers slightly better performance in the United States than in Europe.

There are a few drawbacks to this approach. First, the information contained in the middle quantiles is wasted, because only the top and bottom quantiles are used in forming the hedged portfolio. Second, it is implicitly assumed that the relationship between the factor and future stock returns is linear or at least monotonic (they move in the same direction but not necessarily at a constant rate), which may not always be the case. Third, equally weighting all stocks in the long and the short quantiles does not properly take account of each stock's volatility and its correlation with other stocks in the portfolio. Fourth, the hedged portfolio approach requires managers to short stocks. Shorting may not be possible in some markets or may be overly expensive in others, particularly some emerging markets. Notably, transaction costs are not incorporated in the backtesting.

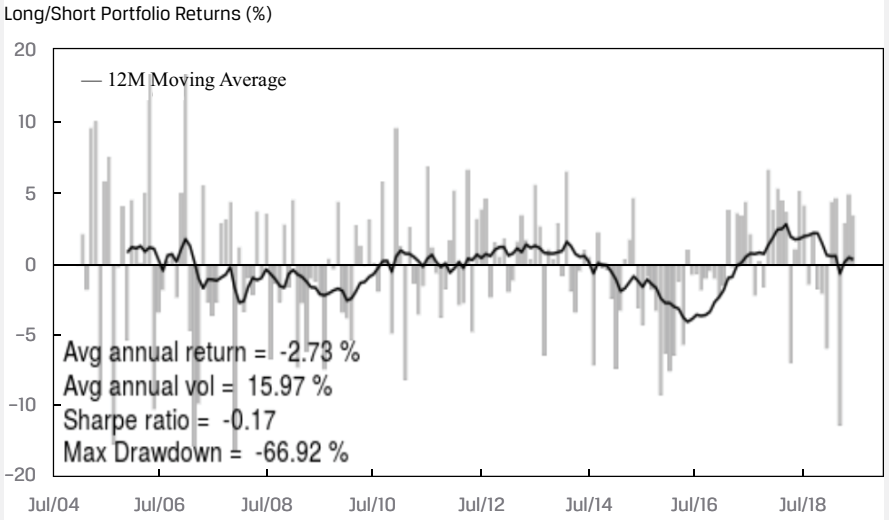
EXAMPLE 2

Performance of Momentum Factor in Mainland China Equity Market, 2004–2019

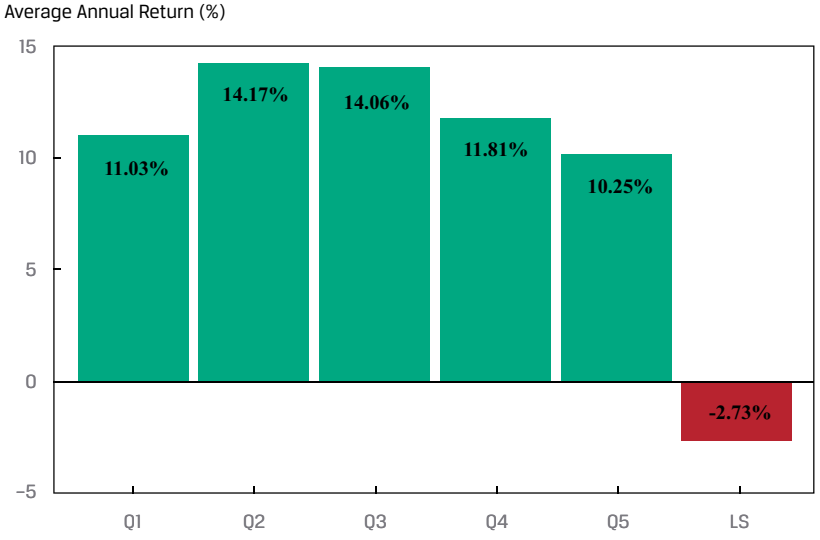
At first glance, Panel A of Exhibit 4 appears to show that the price momentum factor has produced counterintuitive results in Mainland China's domestic equity market during the July 2004–May 2019 period. Backtesting results indicate that the long/short quintile portfolio that invests in the highest momentum bucket and shorts the lowest momentum bucket of stocks delivered a negative return of about –2.7% per annum over the period. However, examining Panel B shows that the payoff pattern is actually an inverse U-shape curve. So, although the highest momentum quintile slightly underperformed the lowest momentum quintile—at 10.3% and 11.0%, respectively—the middle three quintiles substantially outperformed both of the extreme portfolios. There are a few explanations for why the momentum effect is not apparent in Mainland China equities. One explanation, for example, is that the Chinese Mainland equity market consists largely of retail investors (rather than institutional investors) and retail investors tend to overreact to news, which is a behavior that may lead to a mean-reversal effect.

Exhibit 4 Price Momentum Factor, China A-Shares Market

A. Long/Short Hedged Quintile Portfolio Return: 12M-1M Momentum



B. Average Return, by Quintiles: 12M-1M Momentum



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

Describe with respect to Examples 1 and 2 some of the issues with the simple long/short hedged portfolio backtesting approach, and explain why it is important to know your investment universe.

Solution:

The long/short hedged portfolio backtesting approach implicitly assumes the pattern between quantile portfolios and future stock returns is at least monotonic. It also uses information from only the highest and lowest quantiles and thus wastes potentially important information in the middle quantiles. As shown in the example of the price momentum factor in the Mainland Chinese equity market, stocks in the middle quintiles produced higher returns than those in both extreme quintiles. Therefore, for this market, the results from the long/short

hedged portfolio backtesting approach are not so meaningful. So, although the medium-term price momentum effect is well documented in the United States and in many other markets (e.g., Europe), we observe a very different pattern in China A-shares. The bottom line is that depending on the equity market in which one invests, one cannot assume the same anomaly exists and operates in the same way in all markets.

3.5 Pearson and Spearman Rank IC

The information coefficient (IC) is more commonly used than the hedged portfolio approach by practitioners as the standard metric of a factor's predictive power for future stock returns. Since most quantitative models are linear, the IC approach captures the entire spectrum of stocks. This approach contrasts with the long/short quantile portfolio approach, which focuses only on the top and bottom extremes. Therefore, the IC is generally considered to be the better measure for identifying and assessing factors than the quantile-based hedged portfolio approach.

The **Pearson IC** is the simple correlation coefficient between the factor scores (also known as "factor loadings")—for such factors as earnings yield or momentum for the prior period, denoted by f_{t-1}, for all stocks in the investment universe under consideration—and the current period's stock returns, r_t:

$$\text{Pearson IC} = \text{Correlation}\left(f_{t-1}, r_t\right). \tag{3}$$

Because it is a correlation coefficient, the IC's value is always between −100% and +100% (or −1.0 and +1.0). The higher the average IC, the higher the predictive power of the factor for subsequent returns. In practice, any factor with an average monthly IC of 5%–6% is already considered very strong in this context. An important caveat for the Pearson IC, however, is that it is sensitive to outliers.

A similar but more robust measure often preferred by practitioners is the **Spearman Rank IC**. The Spearman Rank IC is essentially the Pearson IC between the prior-period ranked factor scores and the ranked current-period returns:

$$\text{Spearman Rank IC} = \text{Correlation}\left(\text{rank}\left(f_{t-1}\right), \text{rank}\left(r_t\right)\right). \tag{4}$$

Exhibit 5 shows the Pearson IC and Spearman Rank IC for a hypothetical set of nine stocks (A to I) at a point in time (i.e., end of period $t - 1$). As you recall from earlier readings, the correlation coefficient is the ratio of the covariance of two random variables to the product of their standard deviations. The exhibit shows that the Pearson IC for these stocks is marginally negative, at –0.80% (i.e., –0.008), suggesting that the signal did not perform well, and was negatively correlated with the subsequent month's (i.e., period t's) returns. Looking more carefully, however, one can see that the sample factors are generally in line with the subsequent stock returns. The exception is stock I, where the factor predicts the highest return (given that it has the highest score of 1.45), whereas the stock turns out to be the worst performer (–8.50%). This example demonstrates that a single outlier can turn what may actually be a good factor into a bad one because of the Pearson IC's sensitivity to outliers.

Exhibit 5 Pearson IC and Spearman Rank IC

Stock	Factor Score	Subsequent Return	Factor Score Rank	Return Rank
A	(1.45)	(3.00%)	9	8
B	(1.16)	(0.60%)	8	7
C	(0.60)	(0.50%)	7	6

(continued)

Exhibit 5 (Continued)

Stock	Factor Score	Subsequent Return	Factor Score Rank	Return Rank
D	(0.40)	(0.48%)	6	5
E	0.00	1.20%	5	4
F	0.40	3.00%	4	3
G	0.60	3.02%	3	2
H	1.16	3.05%	2	1
I	1.45	(8.50%)	1	9
Mean	0.00	(0.31%)		
Standard Deviation	1.00	3.71%		
Pearson IC		(0.80%)		
Spearman Rank IC				40.00%
Long/Short Tercile Portfolio Return				0.56%

Source: Wolfe Research Luo's QES.

In contrast, the Spearman Rank IC of 40.0% suggests that the factor has strong predictive power for subsequent returns. If three equally weighted portfolios (i.e., tercile portfolios) were constructed, the long basket, which includes stocks I, H, and G (ranked 1, 2, and 3, respectively), would have outperformed the short basket, which includes stocks A, B, and C (ranked 9, 8, and 7, respectively), by 56 bps in absolute return in this period. Therefore, in this case, the Spearman Rank IC is consistent with the long/short portfolio approach, but the Pearson IC is inconsistent.

In the same manner as the long/short hedged portfolio approach, the IC is also computed periodically—for example, monthly—in accordance with the rolling window backtesting and portfolio rebalancing methodology described previously. Typically, investment managers are interested not only in the average IC over time but also in the stability or consistency of the IC. Therefore, managers generally compute the risk-adjusted Spearman IC as a primary performance measure:

$$\text{Risk Adjusted IC} = \frac{\text{Mean(IC)}}{\text{Standard Deviation(IC)}}. \quad \textbf{(5)}$$

EXAMPLE 3

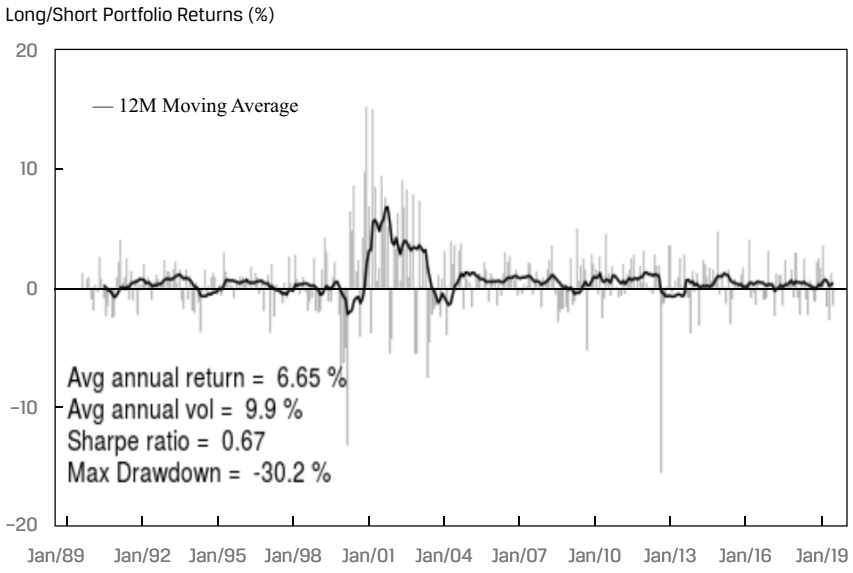
Contrasting Long/Short Hedged Portfolio and Rank IC Approaches in Backtesting

Exhibit 6 compares the backtesting performance of the earnings yield factor in Europe from January 1989 to May 2019 using two different approaches—long/short hedged quintile portfolio (Panel A) versus Spearman Rank IC (Panel B). In Panel A, the bars indicate monthly returns from the long/short hedged quintile portfolio. In Panel B, the bars show the Spearman Rank IC—that is, the rank correlation between the previous month's factor scores and the current month's stock returns.

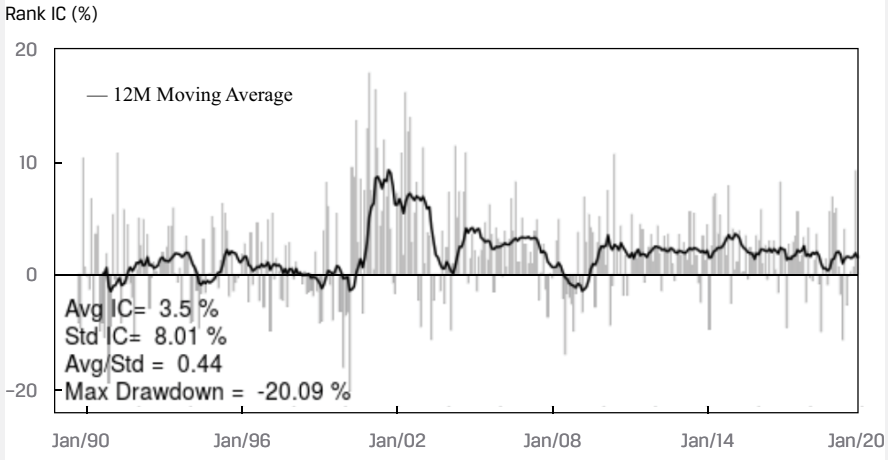
The two evaluation approaches paint a broadly similar picture—namely, that cheaper stocks (ones with higher earnings yields or higher ranked factor scores) have outperformed expensive stocks (ones with lower earnings yields or lower ranked factor scores) in this sample period.

Exhibit 6 Earnings Yield Factor, Performance Comparisons in Europe

A. Long/Short Hedged Quintile Portfolio: Trailing Earnings Yield



B. Spearman Rank IC: Trailing Earnings Yield



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

Describe the pros and cons of using the Spearman Rank IC versus the long/short hedged portfolio approaches to backtesting.

Solution:

The long/short hedged portfolio backtesting approach is more intuitive, since investment managers essentially form paper portfolios and then performance is measured on these hypothetical portfolios. In contrast, the Spearman Rank IC is a measure of the predictive power of a given factor on future stock returns. Importantly, the Rank IC is a correlation, not a portfolio return.

Since the long/short hedged portfolio approach concerns only the top and bottom quantiles of stocks, the middle quantiles are ignored, so information is wasted. In practice, managers typically need to balance risk, return, correlation, transaction costs, trading liquidity, and other issues when they form their portfolios. Since the Rank IC approach captures the predictive power of a given factor via correlation, although it still assumes a linear relationship between the factor score rank and the future stock return rank, it is generally more consistent with actual performance, regardless of how the investment managers construct their portfolios.

3.6 Univariate Regression

Another common way to assess factor performance with backtesting is by performing a cross-sectional (univariate) regression of the following form:

$$r_t = \beta_{0,t} + \beta_{1,t} f_{t-1} + \varepsilon_t, \tag{6}$$

where,

r_t = a vector of stock returns at time t and
f_{t-1} = a vector of stock factor scores at time $t - 1$.

This regression is typically performed monthly and is often referred to as Fama–MacBeth (1973) regression.

The inference centers on whether $\beta_{1,t}$, the fitted factor return, is statistically significant. Note the difference in time subscripts for the factor loading (t) and the factor score ($t - 1$).

In an ordinary least squares (OLS) regression, $\beta_{1,t}$ can be computed as

$$\beta_{1,t} = \frac{\text{cov}(r_t, f_{t-1})}{\text{var}(f_{t-1})} = \frac{\text{corr}(r_t, f_{t-1})\text{std}(r_t)}{\text{std}(f_{t-1})}, \tag{7}$$

where

$\text{cov}(r_t, f_{t-1})$ = the covariance between the prior factor score and the current stock return
$\text{var}(f_{t-1})$ = the variance of the factor scores
$\text{corr}(r_t, f_{t-1})$ = the correlation between the prior factor score and the current stock return—that is, the information coefficient
$\text{std}(r_t)$ = the cross-sectional dispersion (i.e., standard deviation) of stock returns, and
$\text{std}(f_{t-1})$ = the cross-sectional dispersion (i.e., standard deviation) of prior factor scores

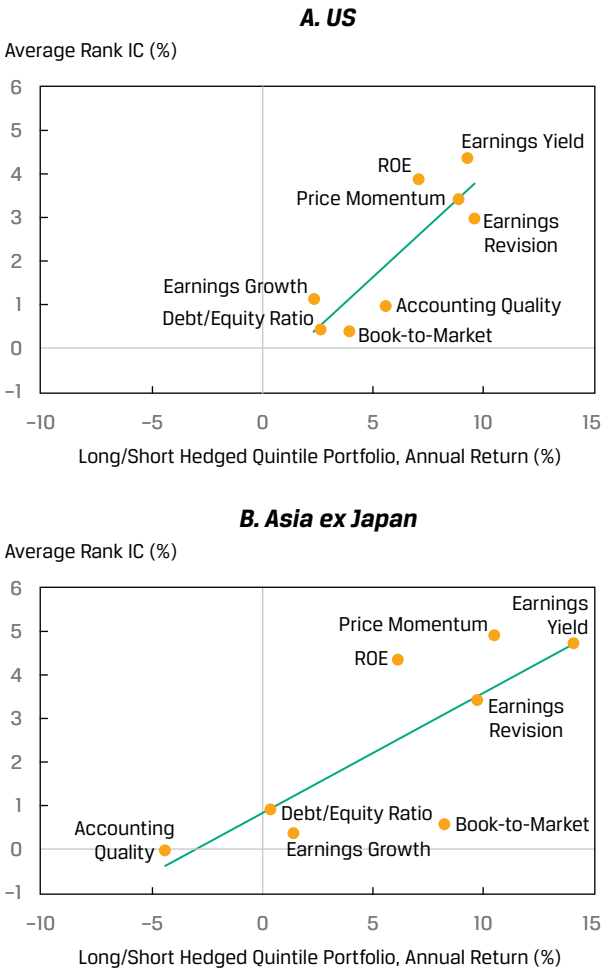
The dispersions in stock returns and factor scores—that is, respectively, $\text{std}(r_t)$ and $\text{std}(f_{t-1})$—are both positive numbers. Therefore, the regression coefficient, $\beta_{1,t}$, and IC, $\text{corr}(r_t, f_{t-1})$, always have the same sign. As a result, the regression approach typically produces results (i.e., whether the results confirm our hypothesis or whether to accept/reject the strategy) similar to those of the IC method.

3.7 Do Different Backtesting Methodologies Tell the Same Story?

Exhibit 7 compares the performance of eight common stock-selection factors in the United States and Asia excluding Japan during 1989–2019 using two backtesting evaluation methods: average Spearman rank IC (y-axis) and long/short hedged quintile portfolio (x-axis). If the factors are arrayed in roughly a straight line, then the two methods yield similar assessments. As shown in Panel A, in the United States, the performance of the eight common stock-selection signals is roughly in line using the two backtesting approaches. In contrast, in Asia ex-Japan (Panel B), we observe some significant deviations. For example, the book-to-market factor appears to be strong using the long/short hedged portfolio approach, with an average annual return of 8.2%. Conversely, based on the Spearman rank IC approach, the book-to-market factor has an average IC close to zero, indicating that it is relatively uncorrelated with future stock returns.

This example illustrates that different backtesting evaluation methodologies do not always tell the same (or even a similar) story. There are many reasons why different approaches can deliver vastly different results. For example, if the true relationship between a factor and future stock returns is highly non-linear, then the Spearman rank IC could be close to zero while the long/short quantile portfolio may still have a highly significant return spread.

Exhibit 7 Long/Short Hedge Portfolio vs. Spearman Rank IC



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

There is no simple rule of thumb as to which approach should be relied on. Ideally, investment strategies should show promising backtested performance using both methodologies. In practice, however, the choice of backtesting methodology depends on a number of considerations, including

- the researcher's personal preference,
- the intended use of the investment strategy, and
- the portfolio construction technique.

For instance, on the one hand, if an investment manager plans to combine factors into a linear multifactor model and use mean–variance optimization to create investable portfolios, then the Spearman rank IC is typically more aligned with the final portfolio performance. When we combine multiple factors into one model, it is not just about the top and bottom quantiles ranked by each factor. Instead, we are interested in the overall distribution of all factors. Furthermore, if mean–variance optimization is used for portfolio construction, we need to rank all stocks and then balance return, risk, trading costs, and liquidity. Again, rank IC is about the overall relationship between a factor and future stock returns, whereas the long/short hedged portfolio approach focuses exclusively on the two extreme quantiles.

On the other hand, if the investment manager intends to construct long/short quintile portfolios for each factor—which is the standard approach used by many alternative beta funds—then that approach should be used as the primary backtesting evaluation method.

EXAMPLE 4

Choosing the Appropriate Backtesting Methodology

A quantitative equity portfolio manager wants to develop a systematic multifactor stock selection model and use a mean–variance optimization technique to construct her portfolio.

Explain whether the manager should use the long/short hedged quantile portfolio approach or the Spearman rank IC approach as the primary decision criterion for evaluating backtests of her portfolio.

Solution:

The investment manager is developing a systematic multifactor model for stock selection and will use mean–variance optimization (MVO) to construct her portfolio. Since she will combine multiple factors into one model, the overall distribution of the factors and the ranking of her stock universe by these factors is crucial. Balancing return, risk, trading costs, and liquidity are also important concerns. For these reasons—and because MVO requires more than just the top and bottom quantiles (ranked by each factor) from the long/short hedged quantile approach—the investment manager should use the Spearman rank IC approach as the primary decision criterion for evaluating backtests of her portfolio.

METRICS AND VISUALS USED IN BACKTESTING

4

c Interpret Metrics and Visuals Reported in a Backtest of an Investment Strategy

In this section, we introduce some of the common metrics and visuals that can assist in interpreting backtesting results and in assessing the effectiveness of a proposed investment strategy.

4.1 Coverage

Not all factors have full coverage in the investment universe. Missing coverage is a practical issue and can be due to a number of reasons—lack of data availability, inapplicability, or outliers. All else equal, managers typically prefer factors with wider coverage.

It is interesting to contrast the price-to-earnings ratio (P/E) with the earnings yield (E/P) factor. The P/E is likely the most commonly used valuation metric among investors; both professionals and retail investors regularly monitor the P/E before making their investment decisions. However, a serious flaw with the P/E is that it cannot be computed if the denominator is zero or negative (i.e., there are no earnings or there are losses), because any such P/E would lack intuitive interpretation. This limitation affects new companies, companies in certain industries (i.e., technology), and companies in distress and turnaround situations, thereby constraining coverage of the P/E metric. Conversely, the earnings yield factor can be computed for any stock, so long as EPS (e.g., positive, zero, or negative) and price data are available.

This difference in coverage between the earnings yield and P/E factors is illustrated in Exhibit 8 for the Russell 3000 from January 1986 to May 2019. Panel A shows P/E coverage averaging about 2,350 stocks, whereas Panel B indicates that the coverage of the earnings yield factor is about 28% higher, at nearly 3,000 stocks. Furthermore, there is a significant difference in the average annual return between these factors. Backtesting of portfolios formed by the long/short hedge method using the P/E (Panel C) and earnings yield (Panel D) factors reveals that the average annual return is 35% higher for portfolios using the earning yield factor (9.24% versus 6.83%). This return difference reflects a strong information signal on the short side. That is, shorting stocks with losses (i.e., negative earnings yield) generated substantial incremental return but also added volatility over this period.

Exhibit 8 P/E Factor vs. Earnings Yield Factor, United States (1986–2019)

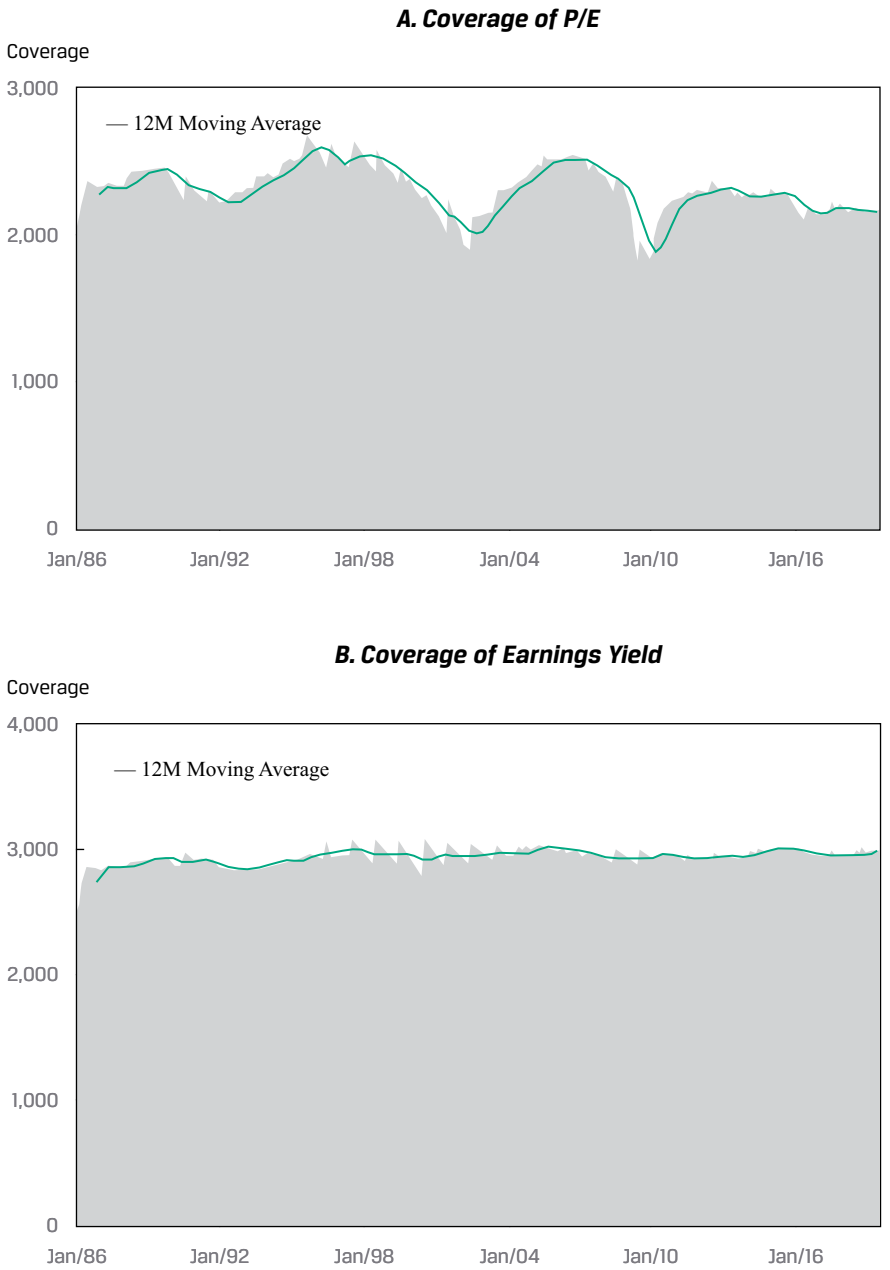
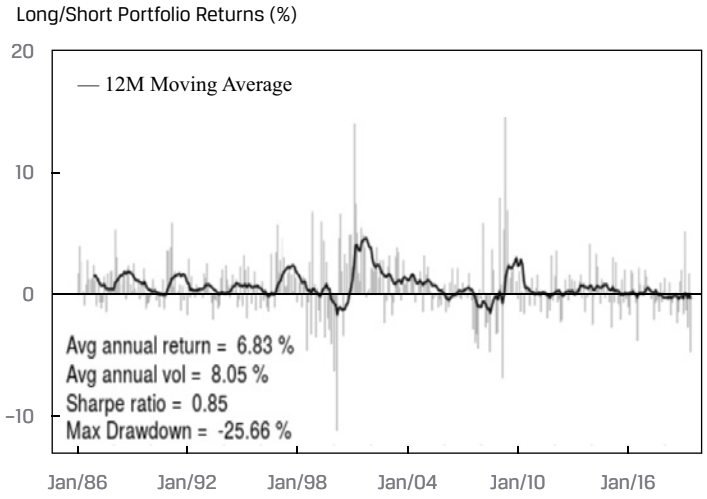
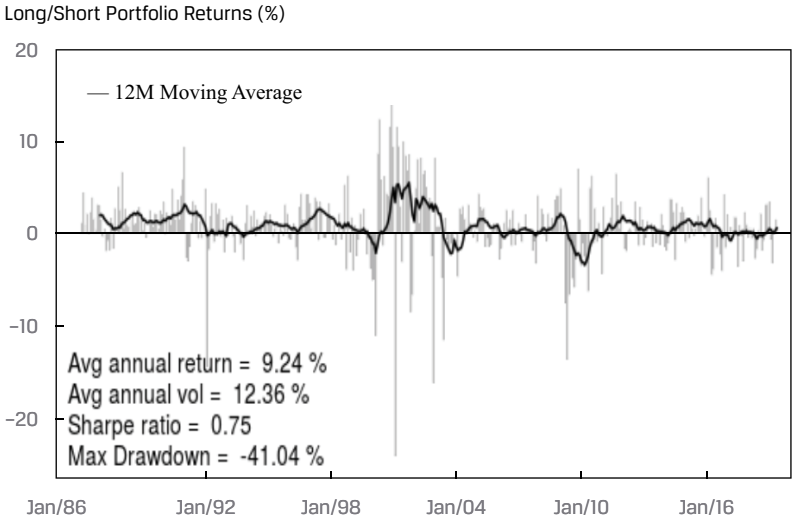


Exhibit 8 (Continued)

C. Performance of P/E



D. Performance of Earnings Yield



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

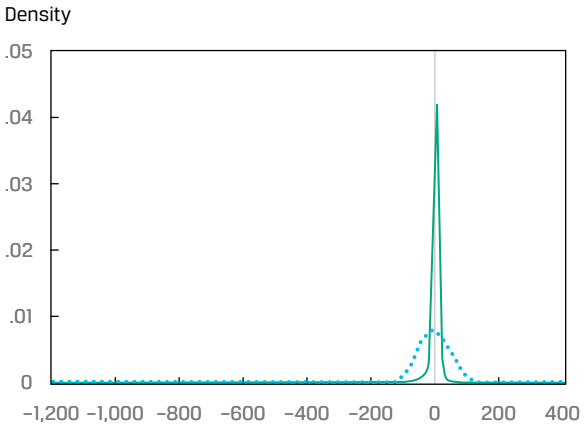
4.2 Distribution

In assessing backtests of multifactor-based models, factor distribution is an important consideration for which two issues need to be addressed. One issue is the factor score distribution. If the distribution of scores is highly skewed with outliers, then the factor scores may need to be transformed. For example, z-score transformation (which re-scales the data into unit standard deviation and centers it with a zero mean) and percentile transformation (which results in a percentile rank factor score) are commonly used in practice. The impact of outliers may be further reduced using winsorization or truncation techniques. Otherwise, combining the skewed factor scores with those of other factors may cause undesirable distortions in the resulting multifactor model. Differences in factor score distributions can be revealed by visual inspection of the probability density function curves of the factors under investigation.

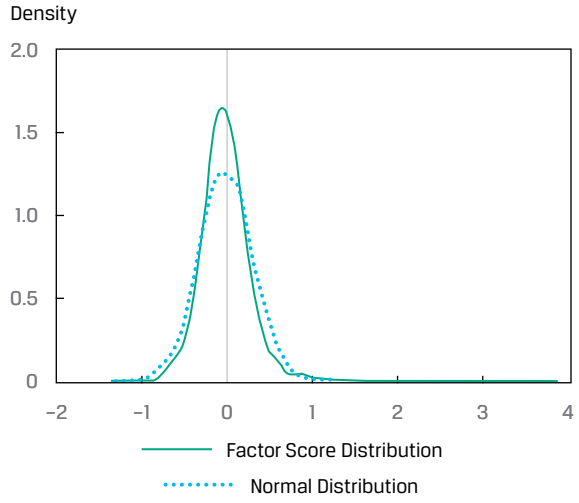
For example, as shown in Panel A of Exhibit 9, compared with a normal distribution (with the same mean and standard deviation), the distribution of raw earnings yield factor scores is clearly skewed to the left, with significant outliers as high as 200% or as negative as –1,000%. The distribution of price momentum factor scores, however, appears to more closely resemble the normal distribution (see Panel B of Exhibit 9); more precisely, it is reasonably symmetric but has fat tails. Note that the *x*-axis represents the total return in the past 12 months excluding the most recent month, which ranges from –90% to 200%.

Exhibit 9 Distribution of Value and Momentum Factors in the United States, May 2019

A. Value (Earnings Yield) Factor Scores



B. Price Momentum (12M–1M Total Return) Factor Scores

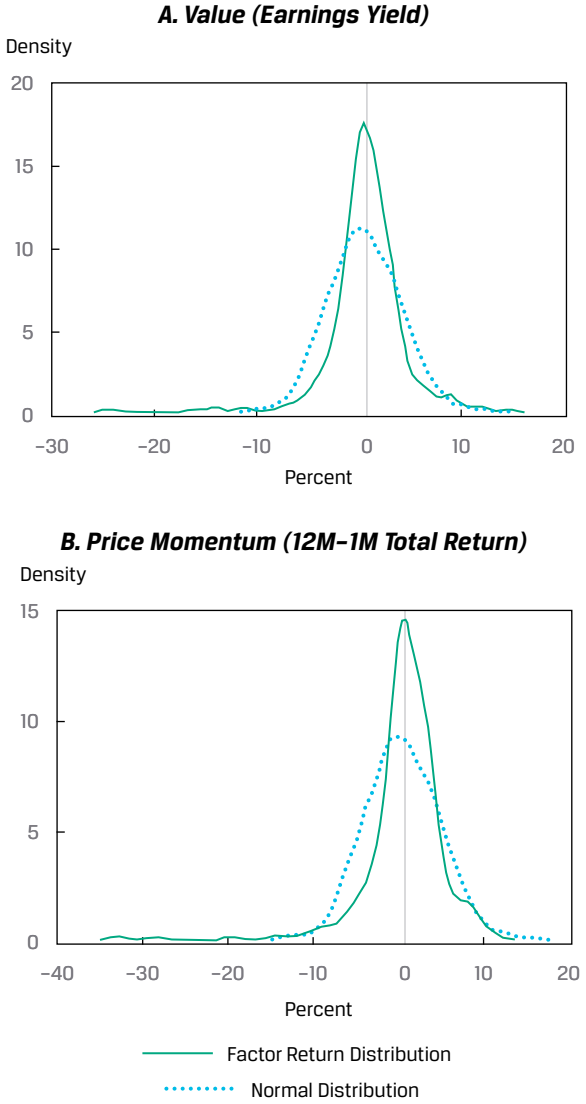


Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES.

In addition to factor score distribution, it is even more important to study the distribution of the investment strategy's returns. The traditional assumption of a normal distribution for asset returns is in fact highly unrealistic in most cases. As shown in Panels A and B of Exhibit 10, the distributions of the classic value and momentum strategy returns, respectively, are clearly non-normal. More problematically, both

factor-based strategies tend to suffer from excess kurtosis (i.e., fat tails) and negative skewness. The excess kurtosis implies that these strategies are more likely to generate surprises, meaning extreme returns, whereas the negative skewness suggests that those surprises are more likely to be negative (than positive).

Exhibit 10 The Distribution of Factor Returns, United States (1986–2019)



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

4.3 Performance Decay, Structural Breaks, and Downside Risk

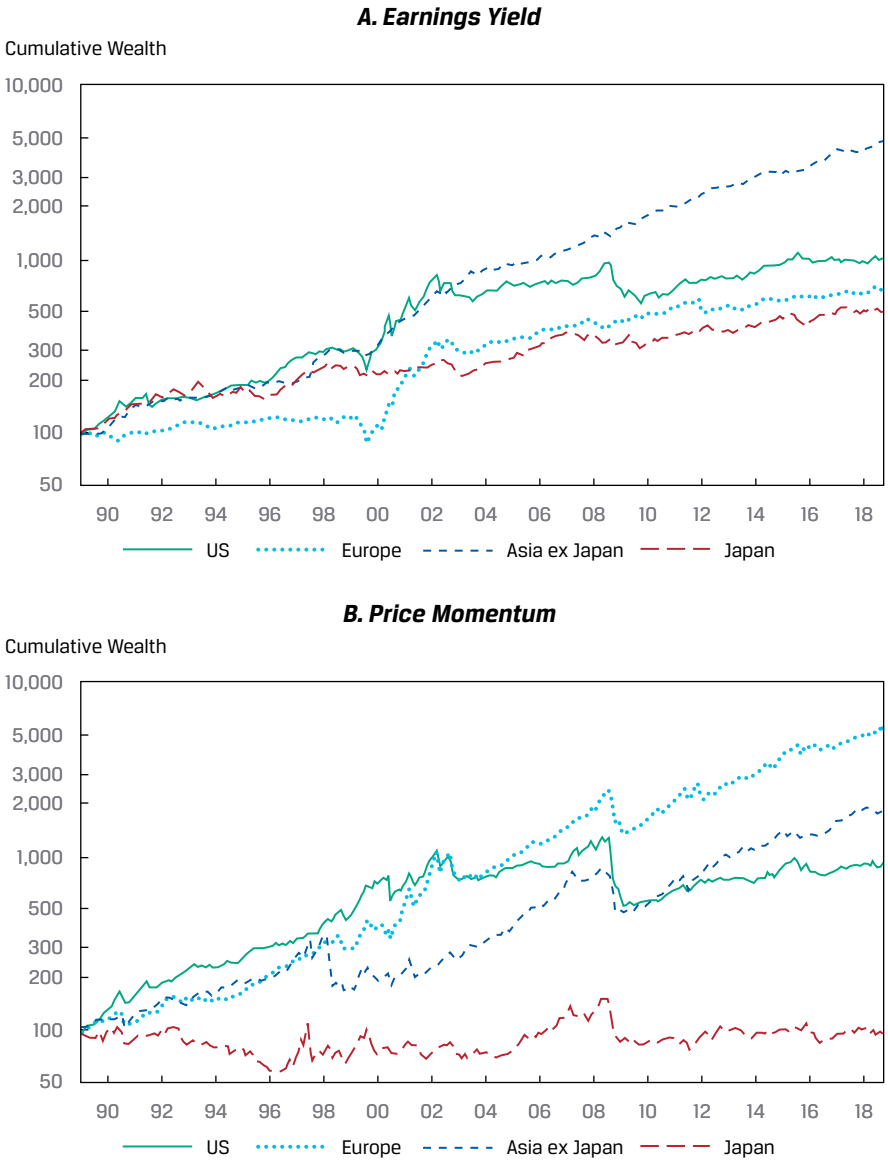
In practice, it is often useful to examine the backtested cumulative performance of an investment strategy over an extended history. For example, the total wealth generated by a long/short hedged quantile portfolio might be calculated, assuming periodic rebalancing (monthly or quarterly). Such performance is often calculated without accounting for transaction costs, because costs depend on the portfolio construction

process, which varies from one manager to another. We recommend plotting performance using a logarithmic scale, wherein equal percentage changes are presented as the same vertical distance on the *y*-axis. Using these cumulative performance graphs, one can readily identify potential performance decay, structural breaks, and downside risk in the backtested investment strategies being assessed.

For example, as shown in Panel A of Exhibit 11, the value strategy (i.e., earnings yield factor) has delivered strong performance over the long run (1990–2019), especially in Asia ex-Japan. However, performance has flattened since 2016 in the United States, Europe, and Japan. Significant drawdowns and potential structural breaks can also be observed in late 1990s (i.e., during the tech bubble) and in March–May 2009 (i.e., the risk rally during the global financial crisis) in most regions.

The price momentum factor has also produced significant excess returns during the same time frame (Panel B of Exhibit 11) in the United States, Europe, and Asia ex-Japan. However, the price momentum effect does not appear to exist in Japan. It is also clear that the momentum strategy suffers from more pronounced periodic downside risk (e.g., March–May 2009) than the value strategy does. The performance of the momentum strategy has also been flattening in the United States since 2016.

Exhibit 11 Cumulative Performance of Earnings Yield and Momentum Factors (January 1990–May 2019)



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

4.4 Factor Turnover and Decay

One of the most common issues concerning interpretation of backtest results of an investment strategy is related to the strategy's turnover. Low factor turnover is desirable, all else equal, since the higher the factor turnover, the higher the portfolio turnover needed to capture the factor. High transaction costs associated with high turnover factors may make them difficult or unrealistic to implement. The turnover of a factor-based portfolio depends not only on the frequency and magnitude of changes in factor scores over time but also on the portfolio construction process itself.

To isolate the effect of factor changes from portfolio construction, signal autocorrelation (i.e., serial correlation) is commonly used to measure factor turnover. Signal autocorrelation is computed as the correlation between the vector of today's (t) factor scores and the factor scores from the preceding period ($t - 1$):

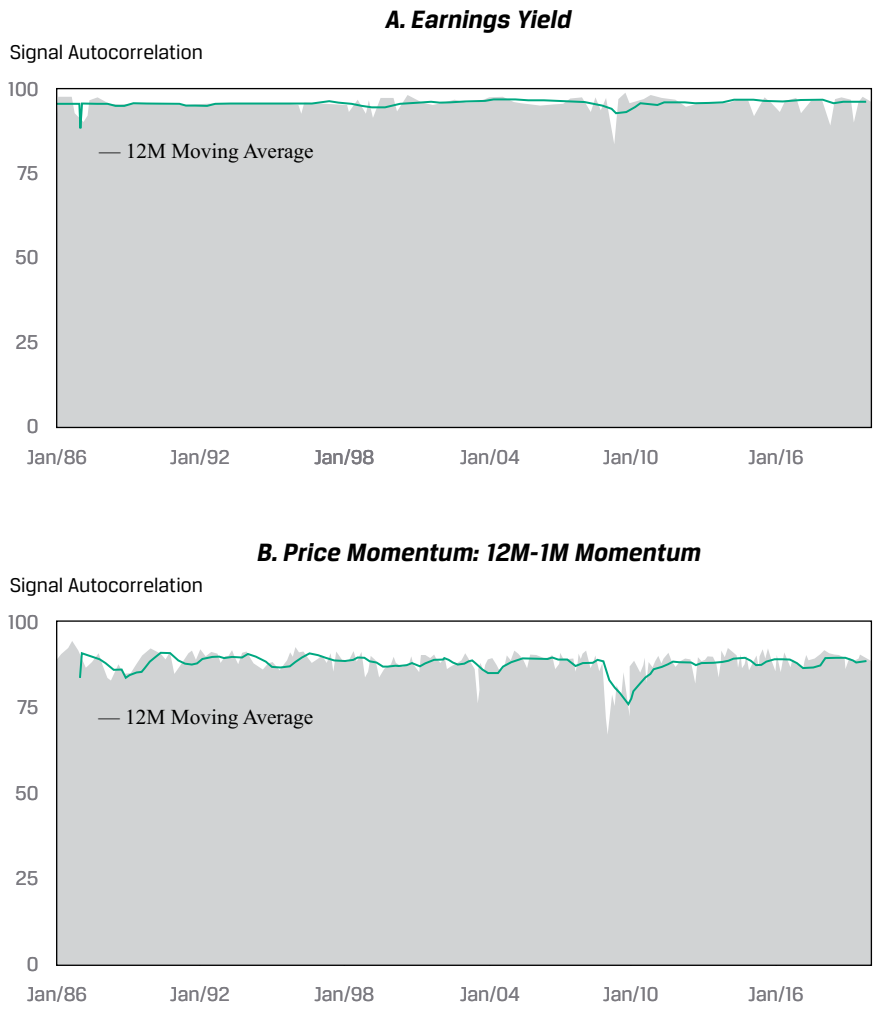
$$\text{Signal autocorrelation}_t = \text{Correlation}\left(\mathrm{f}_t, \mathrm{f}_{t-1}\right), \tag{8}$$

where f_t is the vector of factor scores as of time t.

The signal autocorrelation is then plotted over time. Assuming the identified factors indeed produce strong excess return, all else being equal, factors with low turnover, indicated by high autocorrelation, are preferred because such factors lead to lower portfolio turnover, lower transactions costs, and, therefore, higher after-cost cumulative performance.

For example, as shown in Panel A of Exhibit 12, the average signal autocorrelation for the earnings yield–based value strategy over the period 1986–2019 is 95%. Therefore, because the ranking of stocks based on earnings yield changes very minimally from month to month, the periodic rebalancing of portfolios formed using this factor experience low turnover. Similarly, the average serial correlation for the price momentum factor during that period is about 88% (see Panel B of Exhibit 12), slightly lower than for the value factor, meaning portfolios formed using the momentum strategy experience slightly higher turnover.

Exhibit 12 Signal Autocorrelation: Value vs. Price Momentum Strategies, United States (1986–2019)



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

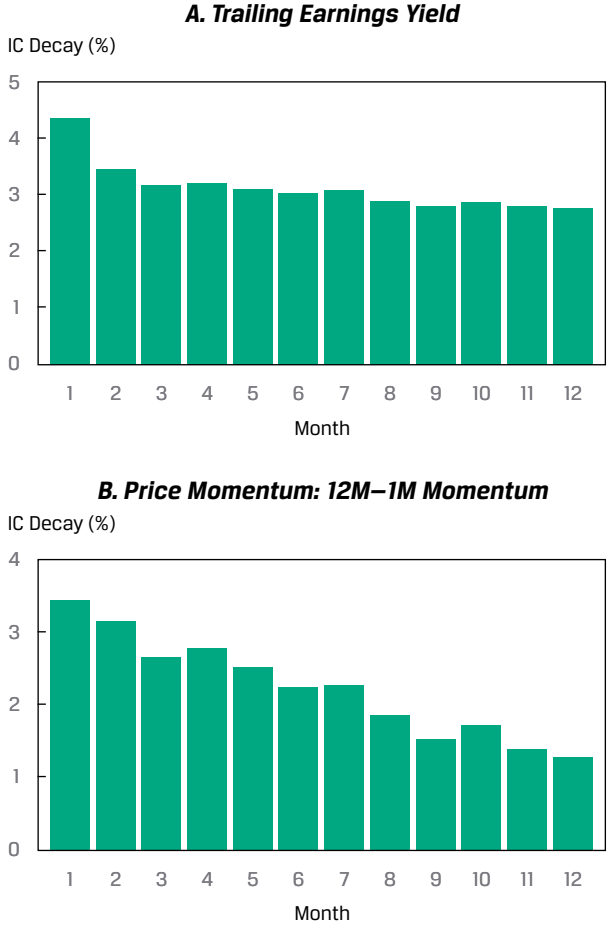
A concept that is closely related but different from factor turnover (i.e., signal autocorrelation) is information decay, which measures the decline in a factor's predictive power as the forecasting horizon (or backtesting history) is extended. Essentially, information decay is proxied by computing the Spearman rank IC between factor scores in the qth month prior (f_{t-q}) and the current month's stock returns (r_t):

$$\text{Spearman rank IC}_q = \text{Correlation}\left[\text{rank}\left(f_{t-q}\right), \text{rank}\left(r_t\right)\right]. \quad (9)$$

We can then plot the IC decay chart to show how long the predictive power of the factors under investigation tends to last. For example, Panel A of Exhibit 13, shows the information decay profile for the US value strategy (based on the earnings yield factor): The predictive power of the underlying factor remains positive (approximately 3% or more) for up to one year. So, the value factor's signal provides strong predictive power, and it decays slowly. In contrast, although the front-month correlation (i.e., IC_1) of the US momentum strategy is only slightly lower than for the value strategy,

the predictive power of the momentum strategy decays much faster (Panel B, with a somewhat different *y*-axis from that in Panel A). In sum, the ideal factor has a high initial IC, a low factor turnover, and a slow factor decay.

Exhibit 13 Signal Decay: Value vs. Price Momentum Strategies, United States



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

5 COMMON PROBLEMS IN BACKTESTING

d Identify Problems in a Backtest of an Investment Strategy

In this section, we discuss some of the most common mistakes investors make when they conduct backtests. Many quantitative investment managers believe that their models are free from human behavioral biases, but we will show how quantitative strategies can suffer from the same biases as fundamental and other investment styles.

5.1 Survivorship Bias

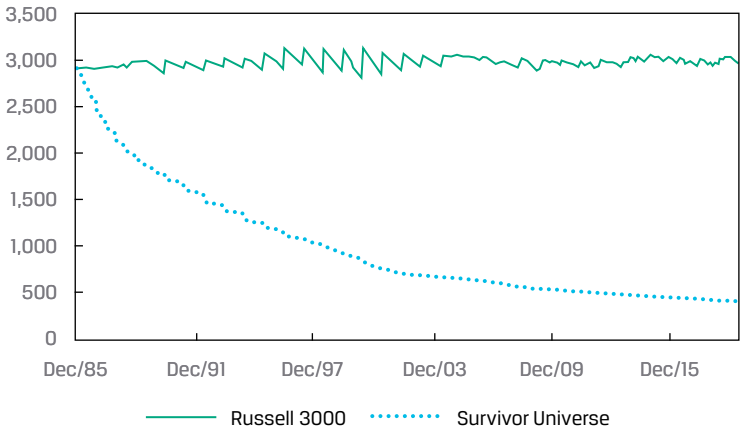
Ignoring **survivorship bias**, the bias that results when data as of a given date reflects only those entities that have survived to that date, is one of the most obvious but, interestingly, also one of the most common mistakes that investors make when conducting backtests. Although widely covered in the academic literature, relatively few practitioners, whether investing in equities, fixed income, indexes, hedge funds, or other asset classes, bother to quantify the real but significant implications of survivorship bias in their backtesting.

Although it is straightforward to backtest an investment strategy with the companies that are currently in the index (i.e., the survivors), tracking all companies that have ever existed in a correct point-in-time fashion (i.e., the casualties as well as the survivors) is actually not so straightforward. **Point-in-time data** means the exact information that was available to market participants as of a given point in time. Point-in-time data allow analysts to use the most complete data for any given time period, thereby enabling the construction (and backtesting) of the most realistic investment strategies.

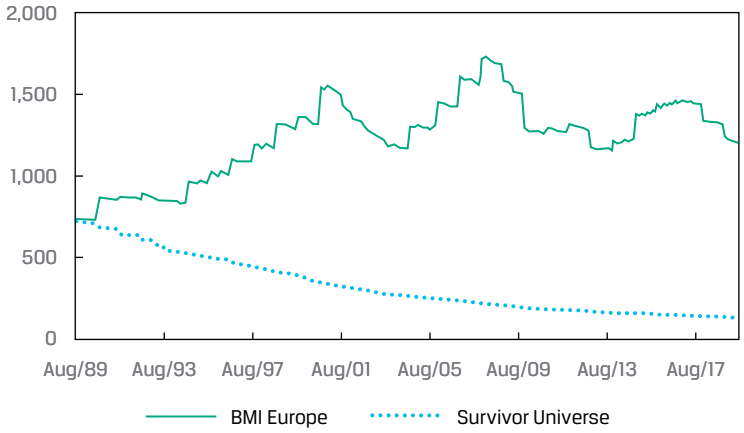
Companies continually appear and disappear. A company can disappear (i.e., be delisted) because of many factors, including privatization, acquisition, bankruptcy, and prolonged underperformance. Similarly, new firms appear via entrepreneurship, spin-offs, and carve-outs that go public and are eventually included in the major indexes. As shown in Panel A of Exhibit 14, the number of true point-in-time companies in the US Russell 3000 Index has stayed relatively stable, at around 3,000, over the past 30 years. However, among the 3,000 companies in the index as of 31 December 1985, less than 400 (or roughly 13%) have survived as of 31 May 2019. Similarly, the S&P BMI Europe Index, which tracks the broad European market, started with about 720 stocks in 1989 and now comprises around 1,200 companies. Among the 720 stocks in the index at inception, only 142 (or about 20%) were still in the index as of May 2019 (Panel B of Exhibit 14).

Exhibit 14 Number of Stocks in Index vs. Survivors

A. US (Russell 3000)



B. Europe (S&P BMI)



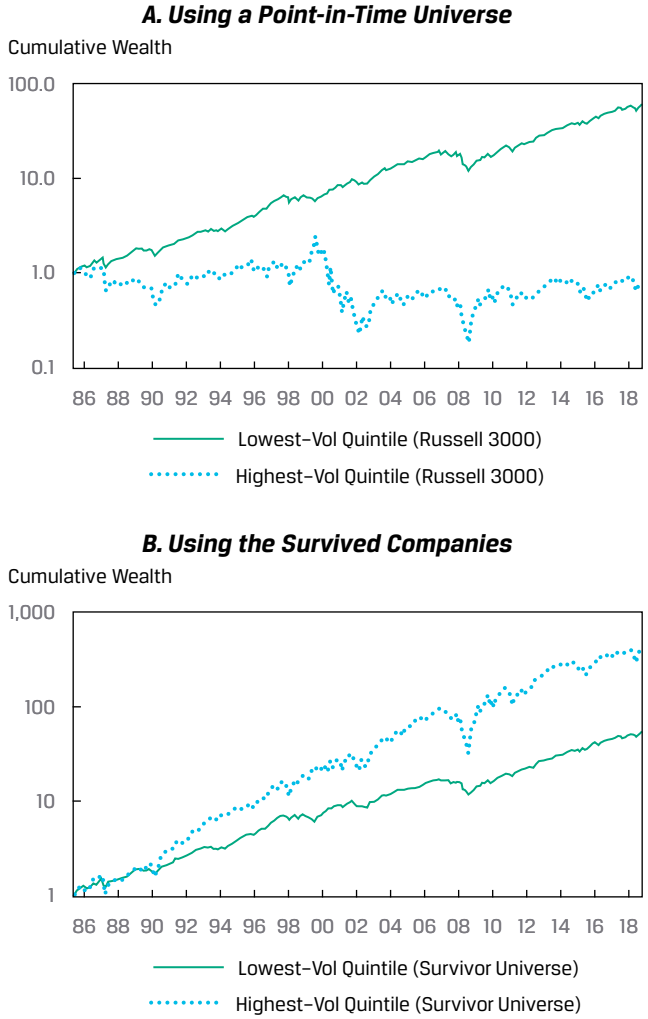
Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES.

Backtesting with only the surviving stocks can create considerable bias and often produces completely incorrect results and conclusions. Unfortunately, tracking companies at each point in time over a long history is not easy, so many practitioners conduct (and data vendors' platforms facilitate) backtesting using the current index constituents. They contend that since you can invest only in the companies that exist today, there is nothing wrong with backtesting strategies on only these firms. However, the problem, as mentioned previously, is that in the past, one could not know which companies would survive in the future, which companies would disappear, and which companies would be created and become successful enough to be added to the index in the future. In sum, backtesting with just current index constituents is a bad practice that can result in faulty investment conclusions.

The danger in ignoring survivorship bias is illustrated using the low-volatility anomaly, a popular investment strategy that argues that stocks with low volatilities tend to outperform high-volatility stocks in the long term. As shown in Panel A of Exhibit 15—here the portfolio is constructed by going long (shorting) the lowest-volatility (highest-volatility) quintile of stocks—a proper backtesting methodology

using the point-in-time Russell 3000 universe finds that low-volatility stocks do indeed very significantly outperform high-volatility stocks over the three decades up to 2019. Importantly, however, if we repeat the backtesting exercise but form the long/short portfolio using only those companies that have survived until the current period, May 2019, then the result is exactly the opposite: The incorrect conclusion now is that high-volatility stocks outperform low-volatility stocks by about 5.5× (see Panel B of Exhibit 15). This example underscores the importance of accounting for survivorship bias in backtesting by using point-in-time index constituent stocks and not just the current survivors.

Exhibit 15 Survivorship Bias and the Low-Volatility Anomaly



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

EXAMPLE 5

Survivorship Bias and the Low-Volatility Anomaly

Explain the rationale behind the big difference in backtesting performance results for the low-volatility anomaly strategy shown in Exhibit 15 using point-in-time data on index constituents versus current data on index constituents.

Solution:

These contradictory results are fairly straightforward to explain. The proper backtesting methodology uses point-in-time data and thus includes current survivors as well as all the stocks that have dropped out of the index—because of mergers and acquisitions or perhaps financial distress or bankruptcy—along with their poor performance during prior periods. So, when the past casualties are properly taken into account, the low-volatility quintile easily outperforms the high-volatility quintile of stocks.

The improper backtesting methodology uses just the current companies in the index (i.e., survivors only). These survivors likely included many companies that in the past experienced financial distress, did not go bankrupt, and ultimately became turnaround successes with considerable upside performance. Such companies likely propelled the outperformance of the high-volatility quintile versus the low-volatility quintile in this biased sample.

5.2 Look-Ahead Bias

Another common mistake investors make in backtesting is failing to recognize and account for **look-ahead bias**. This is the bias created by using information that was unknown or unavailable during the time periods over which the backtesting is conducted. Look-ahead bias is likely the most common mistake that practitioners make when performing backtesting. Survivorship bias can be considered a special case of look-ahead bias, because the question of whether a stock will survive or be added to an index in the future is unknown during the earlier periods over which the backtesting occurs.

Ideally, in backtesting one should use only point-in-time data. Unfortunately, since not all vendor databases provide point-in-time data, there are several issues that must be addressed.

First, analysts must make reporting lag assumptions. For example, we would not have EPS results for the quarter ending 31 December 2018 for all companies on 31 January 2019, since many companies will not yet have reported their earnings. Therefore, analysts typically compensate by adding several months of reporting lag. However, this process can also introduce stale information. Continuing the example, by 31 January 2019, many companies, especially the larger-cap ones, will have reported earnings, but others, especially mid- and small-cap companies, will not have had the chance to report their Q4 2018 earnings. By using the assumption of a reporting lag of three months, we essentially assume that Q4 2018 data are available for all companies on 31 March 2019. In this case, look-ahead bias is significantly reduced. However, because most companies would have reported earnings before 31 March 2019, by using the three-month lag assumption for backtesting done on 31 January and 28 February 2019, we would be using stale financial data; this typically makes backtesting overly conservative.

Instead, using point-in-time data effectively solves this reporting lag problem—no assumption is needed—since we simply use the best information available at any given point in time for our research. In this case, if by 31 January 2019 a company has reported Q4 2018 earnings results, then we would use them; otherwise, we would use Q3 2018 earnings or whatever was available as of 31 January 2019.

A second problem is that companies often re-state their financial statements owing to corrections of accounting errors or changes in accounting policies. Economic data from government agencies are also often being re-stated. Traditional databases keep only the latest numbers or the last re-stated financial statements. By using such databases, an analyst trying to build realistic investment scenarios going back in time would be using information that was not available during the earlier periods of the backtesting. Another form of look-ahead bias arises when data vendors add new companies to their databases. When doing so, they often add several years of historical financial statements into the system. Thus, an analyst backtesting with the current database would be using companies that were not actually in the database during the backtesting period. The consequence of this look-ahead bias is often overly optimistic results.

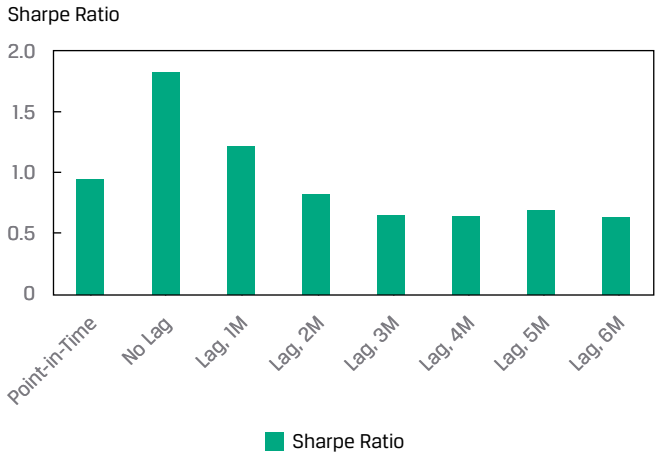
The third problem stemming from traditional non-point-in-time databases is survivorship bias. As mentioned previously, because of merger and acquisition activities, bankruptcy, delisting, and other forms of corporate actions, corporate stocks are constantly being removed from these databases.

To demonstrate the impact of look-ahead bias and the reporting lag assumption, we conduct monthly backtesting using the earnings yield factor. The benchmark is a proper point-in-time database with the actual EPS data as of each month end, so it is free from look-ahead bias. Next, we perform backtesting using EPS data without any reporting lag; this assumes EPS data become available immediately after the close of the quarter (or any other reporting period), so it suffers from full look-ahead bias. Lastly, we add a series of reporting lags, from one to six months.

As shown in Panel A Exhibit 16, it is clear from the backtesting results of the point-in-time and no-lag scenarios that look-ahead bias inflates the performance of our value factor in the United States by almost 100%. The impact of look-ahead bias is evident in all regions. In the United States, Canada, and Japan (Panel B), it appears that a reporting lag of between one and two months produces backtest results that are consistent with those of the proper point-in-time data. Note that using a reporting lag beyond two months introduces stale information and drags performance down significantly. In Europe, the United Kingdom, and Australia and New Zealand, or ANZ (Panel C), a lag assumption of between two and three months appears appropriate, whereas for Asia ex-Japan (AxJ), Latin America (LATAM), and emerging Europe, Middle East, and Africa, or EMEA (Panel D), the point-in-time consistent lag assumption increases to three months. These different lag assumptions reflect the timeliness with which companies in each region report their earnings.

Exhibit 16 Look-Ahead Bias: Impact on Backtesting of Reporting Lag Assumptions (1986–2016)

A. US



B. US, Canada, and Japan

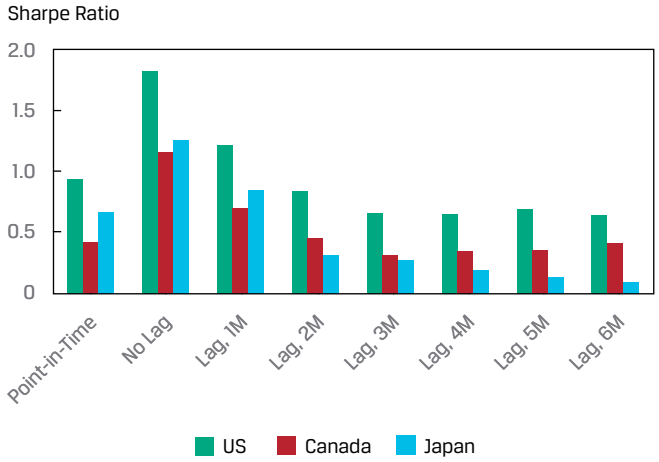
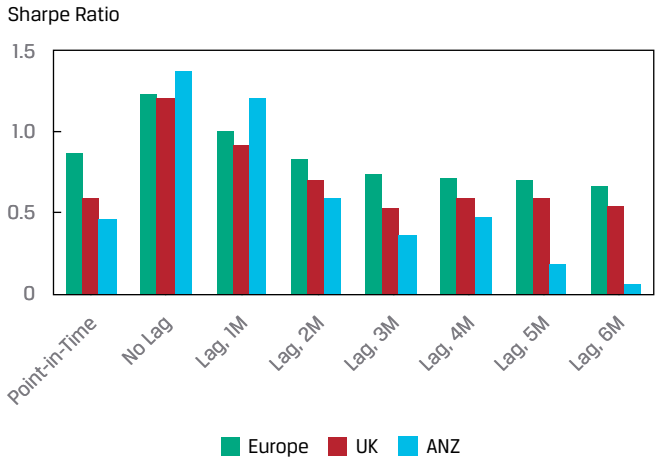
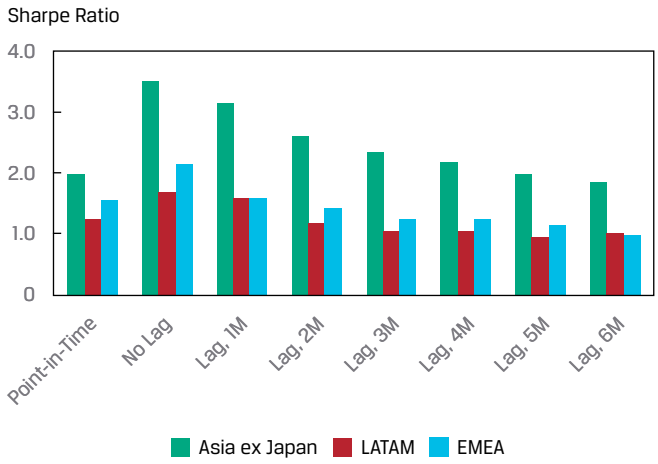


Exhibit 16 (Continued)

C. Europe, UK, and ANZ



D. Asia ex Japan, LATAM, and EMEA



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

BACKTESTING FACTOR ALLOCATION STRATEGIES

6

e Describe Different Ways to Construct Multifactor Models

Few investment managers use a single signal in their models. In practice, most stock selection models share some common multifactor structure, with a linear combination of factors being the dominant framework. Similarly, many fundamental managers use some form of stock screening, and most such filtering systems use more than one factor. In this section, we use a benchmark factor portfolio, which equally weights factors, and a **risk parity** factor portfolio, which weights factors based on equal risk contribution, to discuss how to combine factor portfolios in a multifactor allocation framework. We focus on the benchmark and risk parity factor portfolios since their factor weighting schemes—equal weights and equal risk weights, respectively—are objective and unambiguous.

6.1 Setting the Scene

For demonstration purposes, we choose a few common factors from each main investment style (i.e., value, growth, price momentum, analyst sentiment, and quality):

1. Defensive value: Trailing earnings yield—companies with high earnings yield are preferred.
2. Cyclical value: Book-to-market ratio—companies with high book-to-market ratios (i.e., cheap stock valuations) are bought.
3. Growth: Consensus FY1/FY0 EPS growth—companies with high expected earnings growth are preferred.
4. Price momentum: 12M total return excluding the most recent month—companies with positive price momentum are preferred.
5. Analyst sentiment: 3M EPS revision—companies with positive earnings revisions are bought.
6. Profitability: Return on equity (ROE)—companies with high ROEs are bought.
7. Leverage: Debt/equity ratio—companies with low financial leverage are preferred.
8. Earnings quality: Non-cash earnings—companies with low accruals are bought. Research suggests that net income with low levels of non-cash items (i.e., accruals) is less likely to be manipulated.

For each factor, we form a portfolio by buying the top 20% of stocks and shorting the bottom 20% of stocks ranked by the factor. Stocks in both long and short buckets are equally weighted. The eight different factor portfolios are each rebalanced monthly. For illustration purposes, we do not account for transaction costs or other portfolio implementation constraints.

A straightforward way to combine these factor portfolios is by equally weighting them. In this section, we call the equally weighted multifactor portfolio the benchmark (BM) portfolio. Researchers have found that such an equally weighted portfolio either outperforms or performs in line with portfolios constructed using more sophisticated optimization techniques (e.g., DeMiguel, Garlappi, and Uppal 2007).

Risk parity is a popular alternative portfolio construction technique used in the asset allocation space. Risk parity accounts for the volatility of each factor and the correlations of returns among all factors to be combined in the portfolio. The objective is for each factor to make an equal (hence "parity") risk contribution to the overall (or targeted) risk of the portfolio. Thus, a risk parity (RP) multifactor portfolio can be created by equally weighting the risk contribution of each of the eight factors mentioned above.

6.2 Backtesting the Benchmark and Risk Parity Strategies

Backtesting an asset allocation/multifactor strategy is similar to the method introduced earlier but has a few more complications, since the rolling window procedure is implemented twice.

First, we form eight factor portfolios at each given point in time (i.e., monthly) from 1988 until May 2019 using the rolling window procedure discussed previously. Once the underlying assets (i.e., factor portfolios) are created, we combine them into multifactor portfolios using the two approaches—equally weighting all factors (i.e., benchmark, or BM, allocation) and equally risk weighting all factors (i.e., risk parity, or RP, allocation).

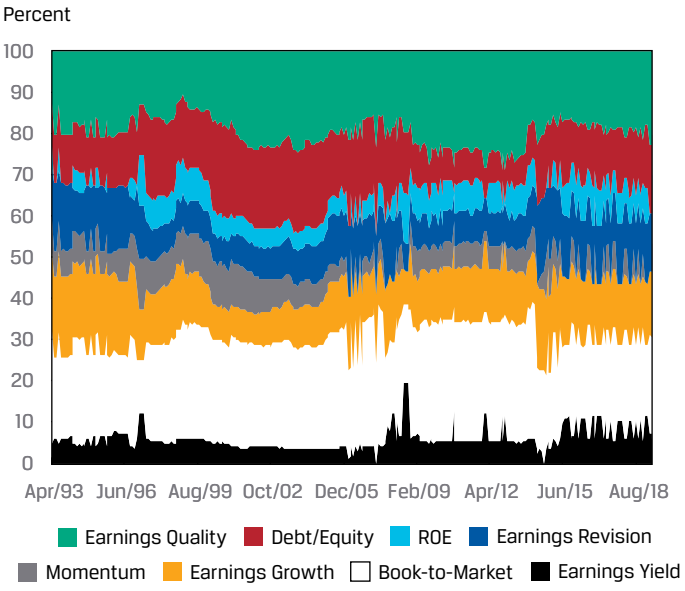
The process for creating the multifactor portfolios requires a second rolling-window procedure, similar to the one presented earlier in Exhibit 2, to avoid look-ahead bias; note that this second rolling window covers the same time span as the first one (i.e., 1988 until May 2019). At each month end, the previous five years of monthly data are used to estimate the variance–covariance matrix for the eight factor portfolios; this is the most important ingredient to form the RP portfolio. Once the covariance matrix is estimated, we can optimize and compute the weights for each of the eight factor portfolios and then form the RP portfolio. Finally, we can compute the returns of the two combination portfolios (BM and RP) during this out-of-sample period using the weights at the end of the previous month and the returns of the eight underlying factors for the current month. This process is repeated every month over the entire horizon of 1988 until May 2019.

We backtested our multifactor strategies using both the equal weighting (benchmark, or BM) scheme and risk parity (RP) scheme for each of the following markets: the United States, Canada, LATAM, Europe, the United Kingdom, emerging EMEA, AxJ, Japan, ANZ, and mainland China. Both multifactor portfolios are rebalanced monthly to maintain equal factor weights or equal factor risk contributions (i.e., risk parity). As noted previously, the key input to the RP allocation is the monthly variance–covariance matrix for the eight underlying factor portfolios derived from the rolling (five-year) window procedure. To be clear, each of the eight factor portfolios is a long/short portfolio. However, our factor allocation strategies to form the BM and RP multifactor portfolios are long only, meaning the weights allocated to each factor portfolio are restricted to be non-negative. Therefore, factor weights for the BM and RP portfolios are positive and add up to 100%.

Panel A of Exhibit 17 shows that the weights of the eight factor portfolios in the RP allocation are relatively stable over time (1993–2019) in the United States. Notably, book-to-market and earnings quality factor portfolios receive the largest allocations, whereas ROE and price momentum factor portfolios have the lowest weights. Although the RP portfolio appears to deliver a lower cumulative return than the BM portfolio does (Panel B), Panel C shows that the RP portfolio's volatility is less than half the volatility of the BM portfolio. As a result, the Sharpe ratio of the RP portfolio is nearly twice that of the BM portfolio (Panel D).

Exhibit 17 Backtesting Multifactor Strategies: Equally Weighted Benchmark Portfolio vs. Risk Parity Weighted Portfolio

A. RP Portfolio Allocation Weights in the US



B. Cumulative Return

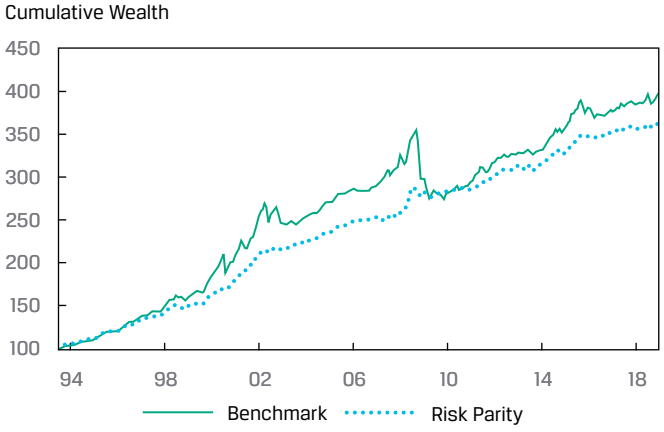
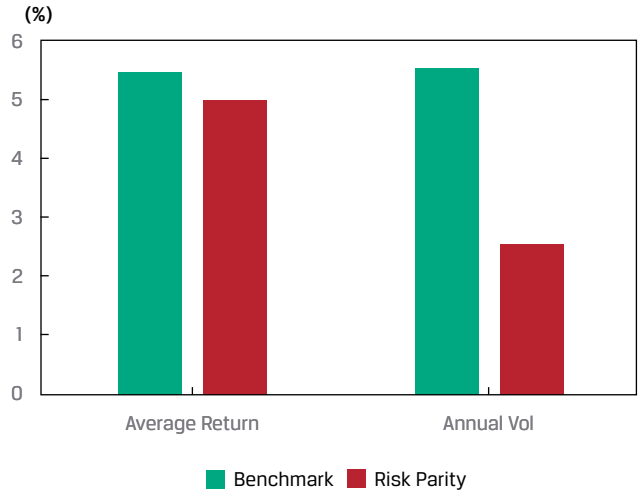
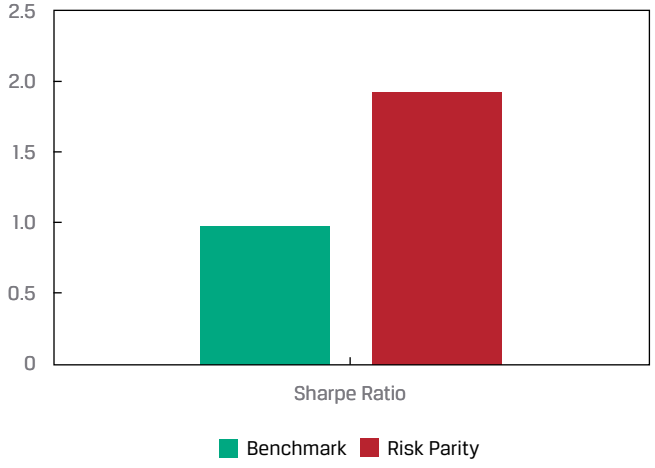


Exhibit 17 (Continued)

C. Average Return and Volatility



D. Sharpe Ratio



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

EXAMPLE 6

Backtesting the Performance of Factor Allocation Strategies

Sarah Koh heads the team at Newton Research Pte. responsible for assessing clients' equity strategies using backtesting and simulation techniques. SWF Fund, one of Newton's biggest clients, has asked for an assessment of two factor-based allocation strategies it is considering implementing.

During the presentation of her backtesting results to SWF's investment committee, Koh is asked the following questions:

1 Regarding rolling window backtesting, which one of the following statements is *inaccurate*?

A The data are divided into just two samples.

B The data are divided using a walk-forward framework, where today's out-of-sample data become part of the next period's in-sample data.

C Repeated in-sample training and out-of-sample testing allow managers to revise their models and readjust security positions on the basis of the arrival over time of new information.

2 Which describes a drawback of the long/short hedged portfolio approach for implementing factor-based portfolios?

A The hedged portfolio is formed by going long the top quantile (with the best factor scores) and shorting the bottom quantile (with the worst factor scores).

B Securities must be ranked by the factor being scrutinized and then grouped into quantiles based on their factor scores.

C Because only the top and bottom quantiles are used in forming the hedged portfolio, the information contained in the middle quantiles is wasted.

3 Regarding the Spearman rank IC approach to implementing factor-based portfolios, which one of the following statements is *inaccurate*?

A The Spearman rank IC is essentially the Pearson IC between the prior-period ranked factor scores and the ranked current-period returns.

B Unlike the long/short hedged portfolio approach, the Spearman ranked IC approach captures the entire spectrum of stocks.

C The Spearman rank IC is more sensitive to outliers than the Pearson IC is.

4 Which one of the following is *not* a metric or visual used in assessing backtesting of a factor-based investment strategy?

A Distribution plots of factor returns

B A word cloud of text describing the characteristics of the factor

C Signal autocorrelation as a measure of factor decay

5 Point-in-time data are useful for avoiding the following problems that may affect backtesting *except*:

A insufficient factor coverage.

B survivorship bias.

C look-ahead bias.

6 Regarding the use of rolling window backtesting in assessing factor allocation to a risk parity–based strategy, which statement is correct?

A The procedure is used once for estimating factor returns over the rolling window.

B The procedure is used once for dividing the data into just two samples.

C The procedure is used twice, once for estimating factor returns over the rolling window and a second time for estimating the covariance matrix of factor returns (for deriving risk parity weights) over the rolling window.

Solution to 1:

A is correct, since the statement is inaccurate. B and C are incorrect, because they accurately describe the rolling window backtesting technique.

Solution to 2:

C is correct, since it best describes a drawback of the long/short hedged portfolio approach. A and B are incorrect because they describe the approach itself.

Solution to 3:

C is correct, since the statement is inaccurate. A and B are incorrect, because they accurately describe the Spearman rank IC approach to implementing factor-based portfolios.

Solution to 4:

B is correct, since a word cloud is not a visual used in assessing backtesting of a factor-based investment strategy. A and C are correct, because they are visuals and metrics, respectively, used to assess backtests of factor-based strategies.

Solution to 5:

A is correct, since this choice is wrong. B and C are incorrect, since point-in-time data are useful for avoiding survivorship bias and look-ahead bias in backtesting.

Solution to 6:

C is correct, since the procedure must be used a second time for estimating the covariance matrix of factor returns (for deriving risk parity weights) over the rolling window. A and B are incorrect.

COMPARING METHODS OF MODELING RANDOMNESS

7

f Compare Methods of Modelling Randomness

The backtesting process as described previously preserves the integrity of the time dimension very well. In backtesting, we essentially assume that we can go back in time, develop our investment strategies, and rebalance our portfolios according a prescribed set of rules. Then, we assess the performance of our investment ideas. It is intuitive, because it mimics how investing is done in reality—that is, forming our ideas, testing our strategies, and implementing periodically.

Importantly, however, we implicitly assume that the same pattern is likely to repeat itself over time. Asset allocation decisions, in particular, have traditionally depended heavily on the assumption that asset returns follow a multivariate normal distribution. In reality, however, asset returns often show skewness and excess kurtosis (i.e., fat tails). Therefore, traditional portfolio construction techniques that depend heavily on estimation of the covariance matrix, such as mean–variance optimization and risk parity, may sometimes yield flawed results. Moreover, conventional rolling window backtesting may not fully account for the dynamic nature of financial markets or potentially extreme downside risks (because it cannot account for such dynamic/extreme events that have not yet occurred). We now explore how scenario analysis and simulation can provide a more complete picture of investment strategy performance.

Traditionally, simulation and scenario analysis are used more regularly in asset allocation, risk management, and derivative pricing compared with quantitative equity investing. In the following sections, we describe how such techniques can be implemented by investment managers.

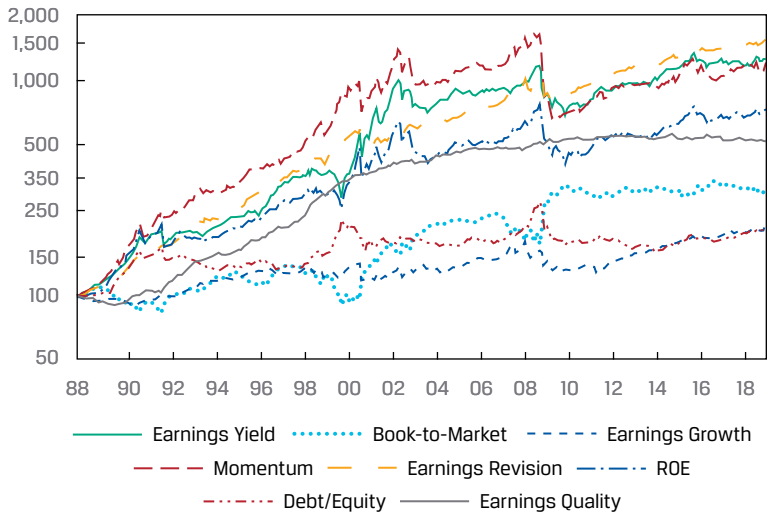
7.1 Factor Portfolios and BM and RP Allocation Strategies

We pick up from the prior discussion, which described the set of eight factor portfolios. To reiterate, for each factor, a portfolio is formed by buying the top 20% of stocks and by shorting the bottom 20% of stocks ranked by the factor. Stocks in both long and short buckets are equally weighted, and all portfolios are rebalanced monthly. Next, we construct the multifactor portfolios (using these factor portfolios) following the two approaches described earlier—that is, the equally weighted benchmark (BM) and the equally risk weighted risk parity (RP) strategies.

As shown in Exhibit 18 (which uses a logarithm scale on the y-axis), all eight factors have delivered reasonable return performance over the long term (1988–2019). In terms of cumulative return, the earnings revision, earnings yield, and price momentum factors produce the highest returns, and the earnings growth and debt/equity factors lag far behind. The eight factor portfolios appear to share some commonalities. Their returns seem to fall into three clusters: (1) earnings revision, earnings yield, and price momentum; (2) ROE and earnings quality; and (3) book-to-market ratio, earnings growth, and debt/equity. They also show significant dispersions at times.

The issue with financial data is that there is only one realization of a time series. Therefore, we must rely heavily on the statistical distribution of asset returns in our modeling. As we will demonstrate, the common assumptions of a multivariate normal distribution and time-series stationarity are often unrealistic, which highlights the importance of using simulation and scenario analysis to supplement the traditional rolling window backtesting.

Exhibit 18 Cumulative Return of Eight Factor Portfolios, United States (1988–2019)

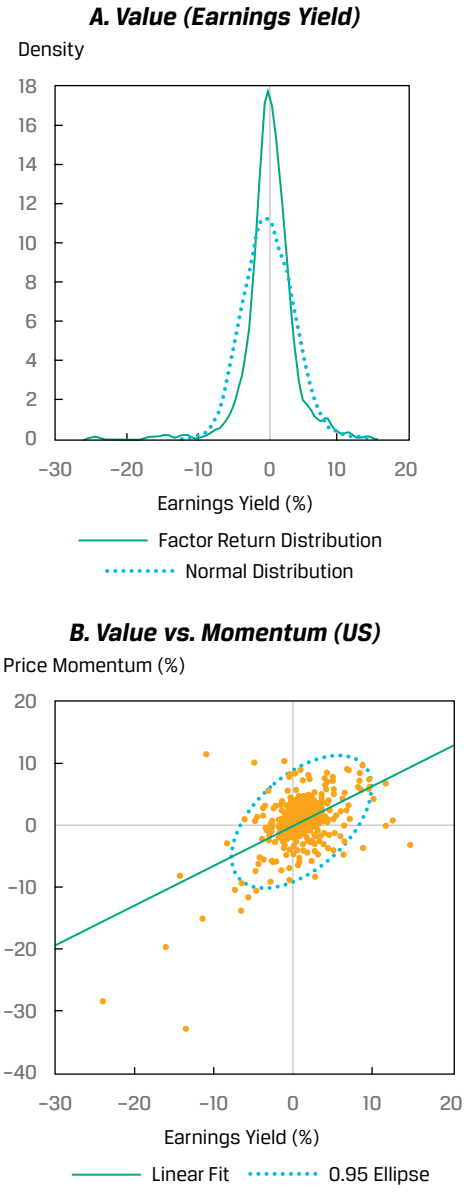


Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES.

7.2 Factor Return Statistical Properties

Asset and factor returns often are left skewed and display excess kurtosis, which is characterized by fat tails, as seen in Panel A of Exhibit 19 for the earnings yield factor. Furthermore, the joint distribution of such returns is rarely multivariate normal, so typically the means and variances of these returns and the correlations between them are not sufficient to describe the joint return distribution. In other words, the return data do not line up tightly around a trend line because of fat tails and outliers. For example, as shown in Panel B, the scatterplot for value (earnings yield) and momentum (12M–1M return) shows some significant deviations from a linear fit, especially at the left tails, where a number of outliers are clearly discernible.

Exhibit 19 Distribution of Selected Factor Returns, United States (1988–2019)



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

7.2.1 *Mean, Standard Deviation, Skewness, and Kurtosis*

If asset and factor returns are normally distributed, then mean and standard deviation should fully capture the randomness in the data. However, the normality assumption is often not valid in investment data, so skewness and kurtosis, the third and fourth moments of the return distribution, respectively, are needed to properly characterize the distribution. As a reminder, skewness measures symmetry (or lack thereof) of the return distribution. Kurtosis measures fat tails (extreme occurrences or outliers)

relative to the normal distribution. For normally distributed data, skewness is zero and kurtosis is three (i.e., excess kurtosis, which is kurtosis minus three, is greater than zero).

Exhibit 20 presents statistics for the return distributions of the eight factor portfolios and the equally weighted BM and RP weighted multifactor portfolios from 1993 to 2019. Six of the eight factor portfolios have negative skewness (the BM portfolio does as well), and all factors and factor allocation portfolios show excess kurtosis (i.e., kurtosis exceeding 3.0. The downside risk (i.e., minimum monthly return) is clearly greater in magnitude than the maximum upside for most factor strategies. The two factor allocation strategy portfolios—BM and RP—both display moderate mean returns (0.5% and 0.4% per month, respectively) and low standard deviations (1.6% and 0.7% per month, respectively) compared with the eight underlying factor portfolios, highlighting the diversification benefits from factor allocation decisions.

Exhibit 20 Monthly Return Distributions: Factor, BM, and RP Portfolios (1993–2019)

	Earnings Yield	Book-to-Market	Earnings Growth	Momentum	Earnings Revision	ROE	Debt/Equity	Earnings Quality	Benchmark	Risk Parity
Mean	0.7%	0.4%	0.2%	0.6%	0.7%	0.5%	0.1%	0.4%	0.5%	0.4%
Median	0.6%	0.1%	0.4%	0.8%	0.8%	0.6%	0.1%	0.4%	0.5%	0.4%
Maximum	14.5%	28.9%	6.2%	11.7%	9.1%	10.8%	11.9%	5.3%	4.3%	3.7%
Minimum	(24.0%)	(12.1%)	(15.8%)	(32.7%)	(18.7%)	(28.0%)	(17.1%)	(2.6%)	(10.9%)	(2.5%)
Std. Dev	3.8%	3.7%	2.1%	4.6%	2.4%	3.9%	2.5%	1.2%	1.6%	0.7%
Skewness	(1.00)	2.82	(2.46%)	(2.36)	(2.39)	(1.92)	(0.58)	0.41	(2.40)	0.51
Kurtosis	11.06	23.61	17.80	16.56	20.76	14.96	11.55	3.87	17.78	5.37

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

EXAMPLE 7

Risk and Return beyond Normal Distribution

Compare return distributions for the BM and RP strategy multifactor portfolios and explain which investment strategy offers the more attractive statistical properties for risk-averse investors (refer to Exhibit 20 to answer this question).

Solution:

The BM and RP portfolios have nearly the same mean monthly returns, at 0.5% and 0.4%, respectively. Although the maximum returns are not much different, the RP factor allocation strategy has a much smaller minimum return (–2.5%) and a significantly lower standard deviation (0.7%) compared with those of the BM factor portfolio (–10.9% and 1.6%, respectively). The RP portfolio is also slightly positively skewed (0.51%) and has moderate kurtosis (5.37), in contrast to the negative skew (–2.40%) and high kurtosis (17.78) of the BM portfolio.

Since the RP portfolio offers similar returns, less downside risk, lower volatility, and slightly higher probability of positive returns (i.e., positive skew) and is less fat tailed (i.e., moderate kurtosis, meaning lower probability of extreme negative surprises) than the BM portfolio is, the RP portfolio has the more attractive distribution properties for risk-averse investors.

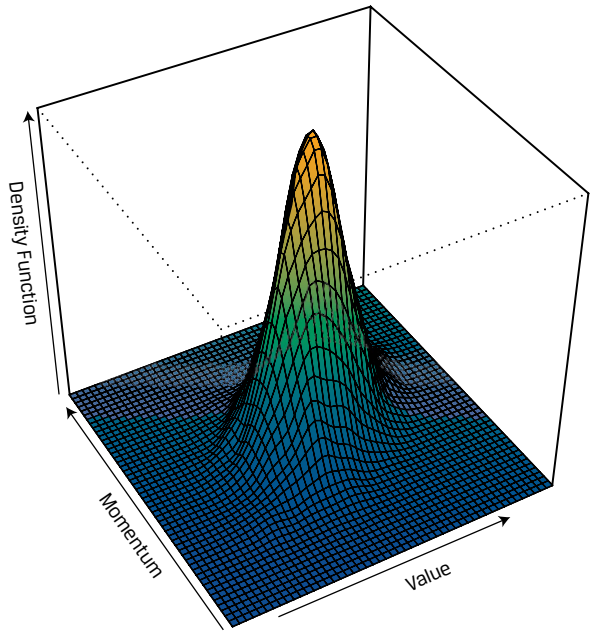
7.2.2 *Tail Dependence*

The **tail dependence coefficient** is similar to the correlation coefficient but focuses on co-movements (i.e., correlation) in the tails of two random variables. Panel A of Exhibit 21 shows the theoretical bivariate normal distribution between the returns of value and momentum factor portfolios over 1988–2019. The theoretical bivariate normal distribution follows a classic bell-shaped curve. However, plotting the empirical distribution between the value and momentum factor returns reveals something quite different. As shown in Panel B, the joint distribution between returns for these two factors is clearly tri-modal, with three distinct peaks (i.e., modes). There is one peak at the center, as would be implied by the bivariate normal distribution. There are also two peaks in the tails, indicating that the probabilities of returns for these two factors move up and down together in the tails. The probabilities in the tails are also much higher than what is implied by the normal distribution. Therefore, this visual (Panel B) confirms that the value and momentum factors have a high, positive tail dependence coefficient.

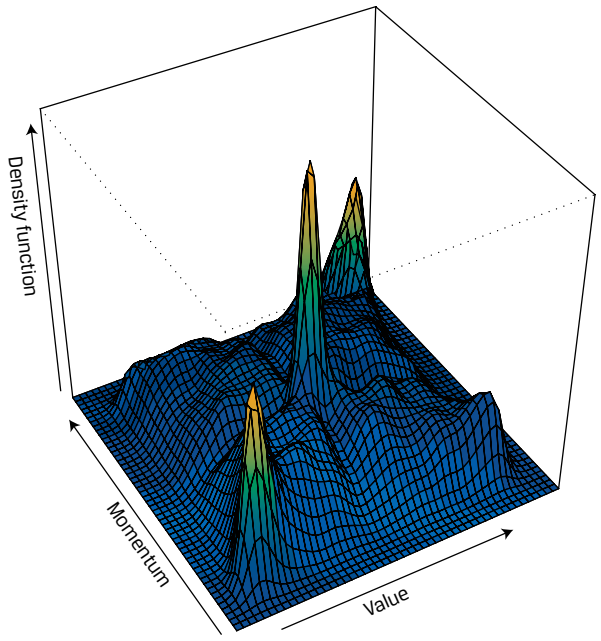
A consequence of failing to properly account for tail dependence (such as that shown in Panel B of Exhibit 21) is that the actual realized downside risk of our portfolios is often higher than what is suggested by our backtesting. The various tools that will be described shortly are designed to help better understand the risk profiles of our strategies.

Exhibit 21 Distributions for Value and Momentum Portfolios (1988–2019)

A. Theoretical Bivariate Normal Distribution



B. Empirical Distribution



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

7.3 Performance Measurement and Downside Risk

The Sharpe ratio and the Sortino ratio (which replaces standard deviation with target semi-deviation in the denominator) are common measures of investment strategy performance. Since we are discussing non-normal return distributions, the focus here is on the downside risk of investment strategies. Given that (negative) skewness, excess kurtosis, and tail dependence are common distributional characteristics of asset (and factor) returns, investment strategies are typically prone to significantly higher downside risk than what is implied by a normal distribution.

7.3.1 *Value at Risk*

Value at risk (VaR) for a given portfolio is a money measure of the minimum value of losses expected during a specified time period at a given level of probability. Although it is widely used to characterize downside risk in terms of the size of the left tail of a portfolio's return distribution, VaR is sensitive to assumptions about the distribution's shape (i.e., fat versus normal tails). Another issue with VaR is that it is not sub-additive, meaning that the VaR of a portfolio can be greater than the sum of the individual risks (i.e., VaRs) of each asset in the portfolio.

7.3.2 *Conditional VaR*

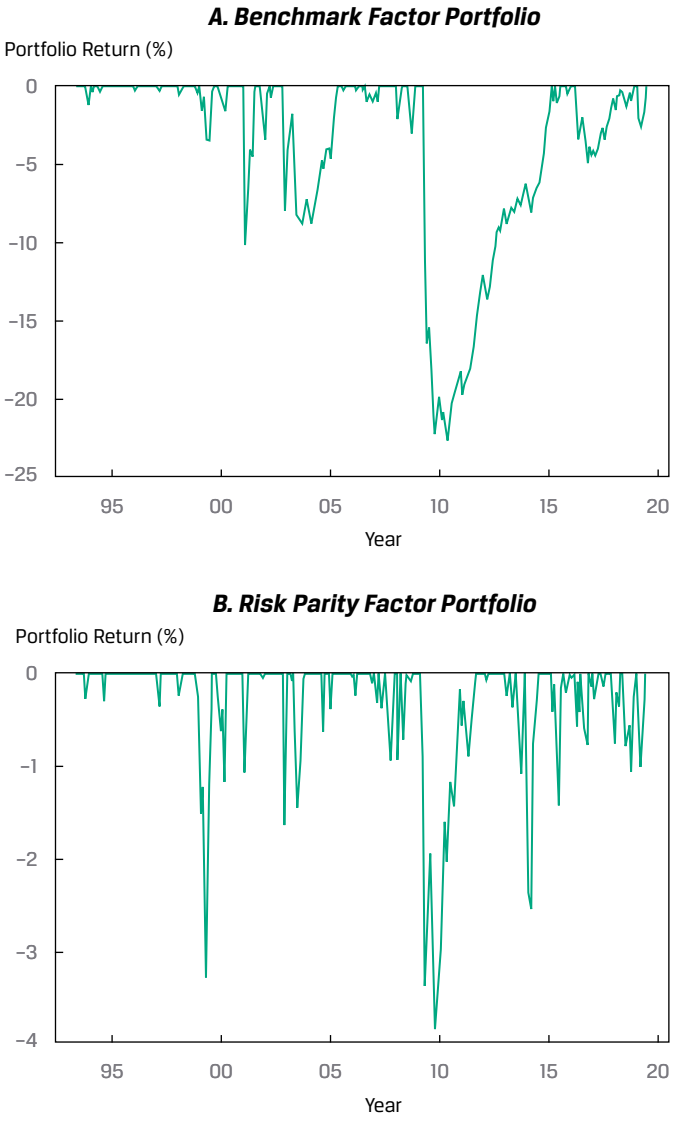
A closely related measure to VaR is **conditional VaR (CVaR)**, which is the weighted average of all loss outcomes in the statistical (i.e., return) distribution that exceed the VaR loss. Thus, CVaR is a more comprehensive measure of tail loss than VaR is. For example, at a preset confidence level denoted α, which typically is set as 1% or 5%, the CVaR of a return series is the expected value of the return when the return is less than its α-quantile. With a sufficiently large dataset, CVaR is typically estimated as the sample average of all returns that are below the α empirical quantile.

7.3.3 *Drawdown Measure*

A widely used measure of downside risk is **maximum drawdown**, the worst cumulative loss ever sustained by an asset or portfolio. More specifically, maximum drawdown is the difference between an asset's or a portfolio's maximum cumulative return and its subsequent lowest cumulative return. Maximum drawdown is a preferred way of expressing downside risk—particularly as associated track records become longer—for investors who believe that observed loss patterns over longer periods of time are the best available proxy for actual exposure. Use of the maximum drawdown measure is particularly common among hedge funds and commodity trading advisers.

Panel A of Exhibit 22, presents the drawdown pattern of the benchmark portfolio, which comprises the eight equally weighted factor portfolios from May 1993 to May 2019. Serious drawdowns, of –8% to –10%, occurred in the early 2000s, but the maximum drawdown for the BM portfolio was more than –20% and coincided with the risk rally in March–May 2009, toward the end of the global financial crisis. Note that downside movements of long/short systematic strategies are mostly associated with market rallies instead of market sell-offs, which is exactly the opposite of long-only market portfolios. As shown in Panel B (which has a different *y*-axis range than Panel A does), the magnitude of drawdowns—between –1% and –4%—for the risk parity portfolio, consisting of the eight equal-risk-weighted factor portfolios, is considerably lower than for the BM portfolio. In fact, the maximum drawdown for the RP portfolio also occurred in the March–May 2009 period and was relatively moderate, at less than –4%.

Exhibit 22 Maximum Drawdown for BM and RP Factor Portfolios (1993–2019)



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

Exhibit 23 compares the various downside risk measures for the eight factor portfolios and the BM and RP portfolios from 1993 to 2019. All three downside risk measures—VaR, CVaR, and maximum drawdown—suggest that the price momentum factor, followed by the ROE factor, has the largest downside risk. The smallest downside risk is observed for the earnings quality factor. As for the factor allocation strategies, the risk parity portfolio shows considerably lower downside risk than any of the eight underlying factors and the benchmark portfolio. This evidence suggests that the RP strategy benefits greatly from risk diversification (in the United States for the period under investigation).

Exhibit 23 Downside Risk Using Monthly Returns: Factor, BM, and RP Portfolios (1993–2019)

	Earnings Yield	Book-to-Market	Earnings Growth	Momentum	Earnings Revision	ROE	Debt/Equity	Earnings Quality	Benchmark	Risk Parity
VaR(95%)	(5.9%)	(0.7%)	(3.9%)	(8.4%)	(3.7%)	(6.8%)	(4.0%)	(1.3%)	(2.6%)	(0.7%)
CVaR(95%)	(14.3%)	(11.1%)	(10.9%)	(22.9%)	(12.8%)	(18.7%)	(8.4%)	(1.7%)	(7.9%)	(0.9%)
Max Drawdown	41.0%	35.3%	27.2%	59.7%	23.9%	47.5%	41.8%	8.3%	22.6%	3.8%

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

7.4 Methods to Account for Randomness

We now discuss several different approaches to account for randomness in asset returns. We will elaborate on the pros and cons of each method and show the differences.

7.4.1 *Rolling Window Backtesting Revisited*

So far, we have used the most conventional approach to simulating the past performance of an investment strategy—rolling window backtesting. The rolling window backtesting technique is widely used among practitioners since it is an intuitive approach that mimics a realistic investment process simulated in the past. The implied assumption with rolling window backtesting is that the future environment will resemble the past, which is often not the case.

Investment time-series data are often non-stationary, with periodic structural breaks. As you may recall, a time series is stationary if its mean, variance, and covariance with itself (for a fixed number of periods) are all constant and finite in all periods. Assuming time-series data to be stationary when in reality they are non-stationary will produce biased (or even invalid) backtesting results. Moreover, other investors are also constantly learning from the data and attempting to improve their own investment decision-making processes; consequently, any alpha from a given investment strategy may be eroded over time. Therefore, the rolling window approach may not necessarily be able to fully capture the randomness in financial data.

Lastly, although rolling window backtesting tests the investment strategy using out-of-sample data, researchers can still easily misuse the technique. For example, financial researchers are often tempted to try various modeling techniques, backtest each of them, and then pick the best performing model. In this case, if the backtested performance does not account for the model selection process, then it suffers from model selection bias. This bias is called **data snooping**, the subconscious or conscious manipulation of data in a way that produces a statistically significant result (i.e., a p-value that is sufficiently small or a t-statistic that is sufficiently large to indicate statistical significance). A preferred approach by leading researchers is, briefly put, to specify an acceptable proportion (q-value) of significant results that can be false positives and establish, on the basis of the q-value and ranked p-values, a critical value for the reported p-values; then, only the tests with smaller p-values are accepted as significant. Alternatively, the data snooping problem may be mitigated by setting a much higher hurdle than typical—for example, a t-statistic greater than 3.0×—for assessing whether a newly discovered factor is indeed adding incremental value (i.e., is statistically significant).

7.4.2 *Cross Validation*

Cross validation is a technique heavily used in the machine learning field. As you may remember, in cross validation, researchers partition their data into training data and testing data (i.e., "validation data"). Essentially, a model is first fitted using the training data, and then its performance is assessed using the testing data. To reduce variability, the process is often repeated multiple times. The validation results are combined over the successive rounds of testing to provide a more accurate estimate of the model's predictive performance.

In machine learning, data sampling is often performed by random draws. For example, in the standard k-fold cross-validation procedure, the original sample data are randomly partitioned into k equal-sized subsamples. At each of the k iterations, a single subsample is held out as the validation data for testing the model and the remaining $k - 1$ subsamples are used to train the model. The process is repeated k times, and the k validation results are averaged to produce a single estimation of the model's predictive performance.

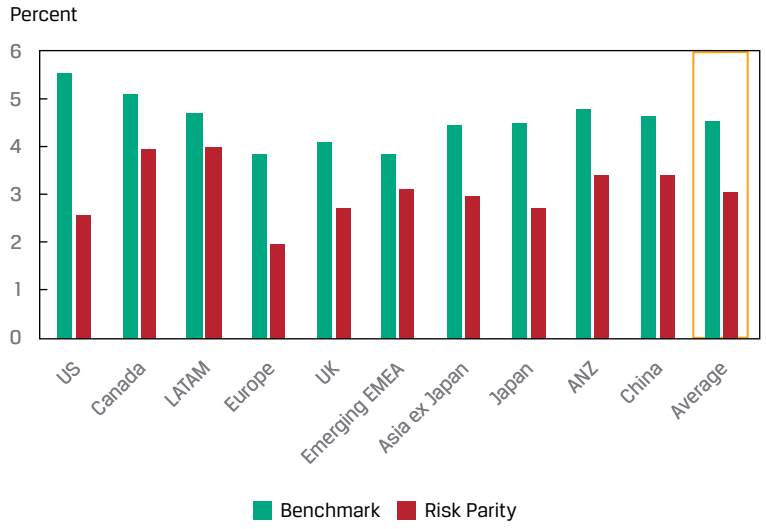
Note that the rolling window backtesting method follows a philosophy similar to that of the cross-validation technique, albeit in a deterministic, non-random manner, meaning that past data (from the rolling window) are used to train a model and then the model is used to invest in the next period.

Another way to perform cross validation is to validate the investment strategy using data from different geographic regions. For example, suppose the risk parity strategy is developed and tested initially using factors based on US equities. The same RP modeling framework can be extended to other markets globally. Then, the average performance from the non-US markets can be used to assess whether risk parity is a robust factor allocation strategy.

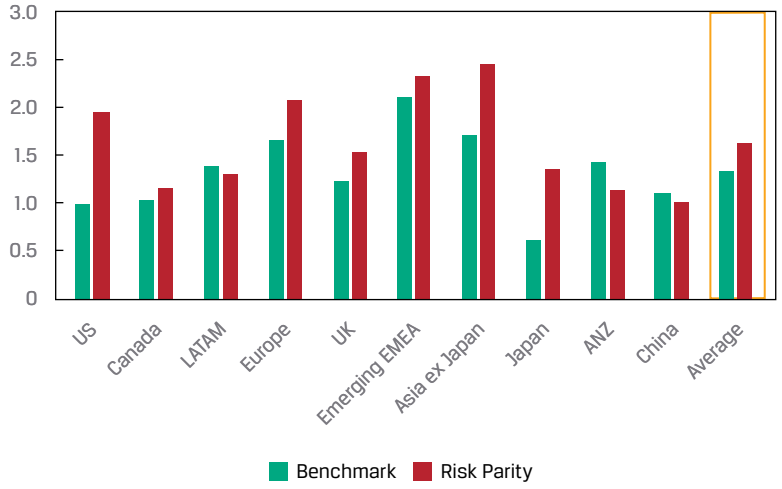
As shown in Panel A of Exhibit 24, as a risk-based factor allocation technique, the RP strategy does indeed deliver a lower realized volatility (i.e., standard deviation of returns) than does the benchmark (i.e., equal-weighted factor) strategy in all 10 global markets over 1993–2019. Similarly, the RP portfolios also outperform the BM portfolios in terms of Sharpe ratio (see Panel B) in 7 of the 10 global markets.

Exhibit 24 Global Cross-Validation, Equally Weighted Benchmark Portfolio vs. Risk Parity Weighted Portfolio (1993–2019)

A. Realized Volatility



B. Sharpe Ratio



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

EXAMPLE 8

Data Snooping in Investment Management

One of the firm's research analysts has just presented to you and several other portfolio managers her risk factor–based quantitative/systematic investment model for the UK market. She reports the development and backtesting of several different models: The number of factors ranged from 5 to 10, rebalancing periods were monthly and quarterly, and rolling windows were implemented for 5, 15, and 25 years of historical data. She recommends the 10-factor model (with monthly rebalancing) since backtesting of 15 years of data generated the

following annualized performance metrics: Sharpe ratio of 3.0 and realized volatility of 1.0%. She also reports a *t*-statistic of 2.5 and a *p*-value of 1.3% for this model of UK market returns, which were the highest and lowest statistics, respectively, of all the models.

Describe the concerns you should raise around the issue of data snooping for this seemingly very attractive strategy.

Solution:

As a portfolio manager, you must be careful in assessing these performance results in light of how the analyst developed and backtested her model. For example, it is critical to know whether backtesting has incorporated transaction costs and trading liquidity. More importantly, however, you need to understand whether data snooping was involved in developing this model/strategy. Given the many variations of models developed and tested by the analyst, it is highly likely that her process suffers from model selection bias. Recommending the model with the highest *t*-statistic and lowest *p*-value also points to data snooping. One way to mitigate the problem is to raise the hurdle for an acceptable model to a *t*-statistic exceeding 3.0 (thereby lowering the *p*-value). Cross validation of the recommended model's results using the *k*-fold technique and testing in other global markets are other techniques that can be used to help the portfolio managers better understand the true performance of this model/strategy.

SCENARIO ANALYSIS

8

g Evaluate and Interpret a Scenario Analysis

A serious issue in backtesting is structural breaks, since backtesting assumes the future will, at least to some extent, resemble history. However, in reality, financial markets often face structural breaks, which can be driven by many exogenous factors. The following are examples of such factors:

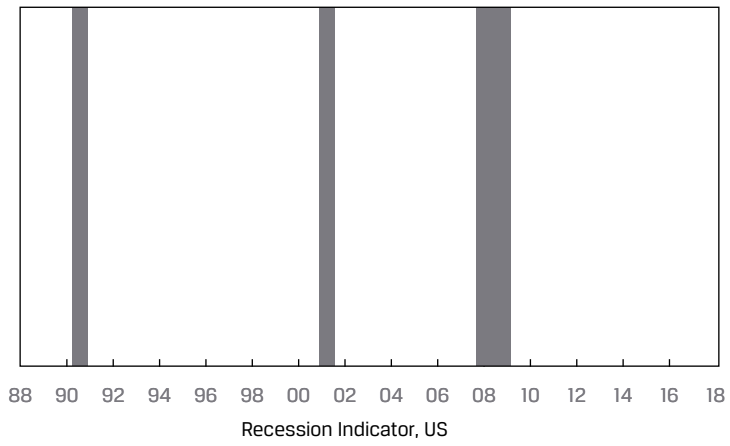
- Geopolitical events, such as changing trade relationships involving countries representing important global equity and bond markets, and exiting or entering major trading blocs by key countries
- Depressions and recessions, such as the 2008–09 global financial crisis
- Major shifts in monetary and fiscal policies, such as the prolonged period of quantitative easing (QE) adopted by major central banks in the aftermath of the global financial crisis
- Major technological changes and advances, such as those that fueled the dot-com bubble and the proliferation of machine learning and artificial intelligence

Importantly, given such structural breaks, potential future outcomes can be highly uncertain. We now demonstrate **scenario analysis**, a technique for exploring the performance and risk of investment strategies in different structural regimes, using two real-world examples:

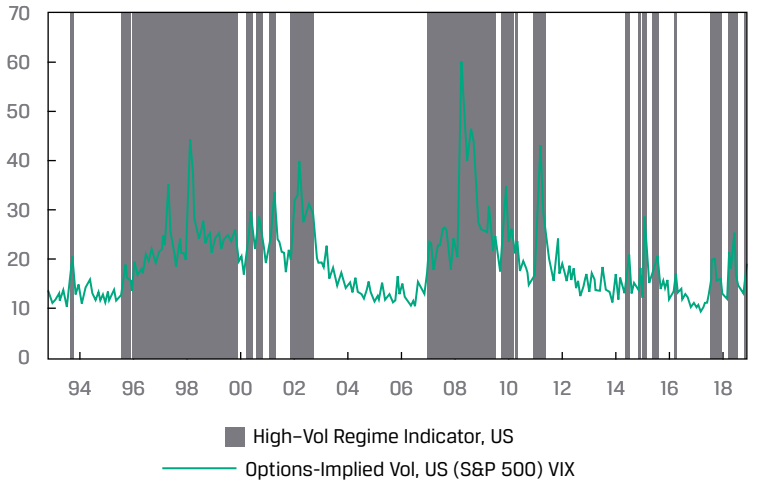
- *Recession environment.* In the United States, since the start of our risk parity allocation strategy in 1993, the National Bureau of Economic Research (NBER) has recognized two official recessions: March 2001–November 2001 and December 2007–June 2009. These recessions are shown in Panel A of Exhibit 25.
- *High- and low-volatility regimes.* The Chicago Board Options Exchange (CBOE) computes the VIX index, which gauges options-implied volatility on the S&P 500 Index. To transform the VIX into a volatility regime indicator, a five-year moving average is computed. Then, the periods when the VIX is above (below) its five-year moving average are defined as high-volatility (low-volatility) regime periods, as shown in Panel B of Exhibit 25 for 1988–2019.

Exhibit 25 Macro-Factor Regime Changes

A. Recession Indicator



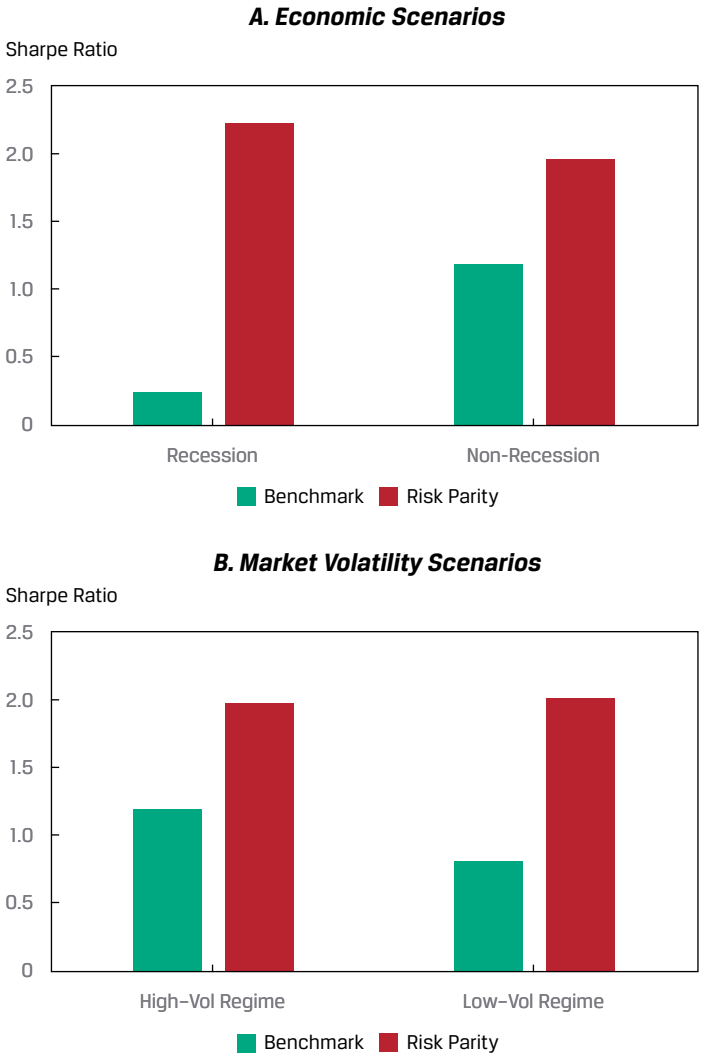
B. VIX: High- vs. Low-Volatility Regimes



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

We can examine the sensitivity of the benchmark and risk parity factor allocation strategies to these two macroeconomic regimes—recession versus non-recession and high volatility versus low volatility. As shown in Panel A of Exhibit 26, in terms of the Sharpe ratio, the RP strategy is quite robust to recession and the BM strategy struggles in recessions. Panel B of Exhibit 26 reveals that the BM strategy's performance is slightly worse in low-volatility regimes than in high-volatility regimes, whereas the RP strategy performs equally well in both volatility environments.

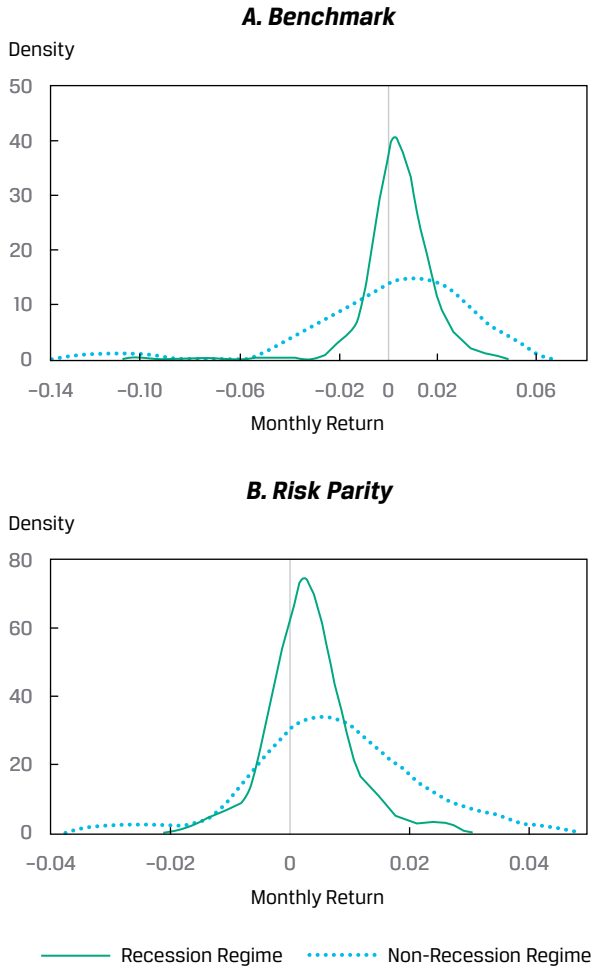
Exhibit 26 Sharpe Ratio for BM and RP Portfolios in Different Macro-Scenarios (1993–2019)



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

In addition to the Sharpe ratio, a density plot can reveal additional information about the sensitivity of the overall distributions of these investment strategies—for example, during recession versus non-recession periods. As shown in Exhibit 27, the distribution of returns for both the BM and RP strategies is flatter in a non-recession environment, which implies higher standard deviations during these regimes. The BM strategy suffers from negative skewness and excess kurtosis (i.e., fat tails to the left), regardless of the recession regime, but its average return is clearly lower in a recession environment (see Panel A). The RP strategy has a lower average return in the recession regime, but its volatility is also much lower (see Panel B); as a result, the Sharpe ratio is about the same in the two regimes (which is consistent with Panel A of Exhibit 26).

Exhibit 27 Distribution of Returns for Factor Allocation Strategies: Recession and Non-Recession Regimes



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

In the previous analysis, the regimes are pre-defined. However, in practice, we do not know the type of regime into which the economy is heading. Although the details are beyond the scope of this reading, it warrants mentioning that a Markov regime-switching model can be used to fit historical data and estimate the probabilities of each different regime in the future. Using these probabilities, we can then compute the weighted average expected return and risk profile of our investment strategy.

A closely related concept is **stress testing**—the process that tests how our strategies would perform under some of the most negative (i.e., adverse) combinations of events and scenarios. For example, an evaluation of the potential downside risks for the BM and RP portfolios could be done via stress testing under the following assumed adverse combination of scenarios: a severe and prolonged economic recession combined with a period of sharply elevated financial market volatility, all at a time when investors have crowded into factor-based investing strategies.

9 HISTORICAL SIMULATION VS. MONTE CARLO SIMULATION

h Contrast Monte Carlo and Historical Simulation

Given that the distribution of asset (and factor) returns may not be multivariate normal, the question is how to account for skewness/excess kurtosis, volatility clustering, and tail dependence. The problem with historical time-series data is that only one set of realized data is observable, and the critical assumption behind classical time-series analysis, that the data are stationary, is unlikely to be true.

Investment data also tend to have high dimensionality. In the factor allocation case, we have eight factor portfolios ($K = 8$), and to construct the risk parity factor allocation strategy, one must estimate the variance–covariance matrix. Assuming a multivariate normal distribution, this requires estimating $[K \times (K + 1)]/2$, or 36, parameters at each monthly rebalancing date. Therefore, even if the assumption of a multivariate normal distribution is reasonable, the uncertainties around the 36 estimated parameters are still difficult to fully account for in a traditional rolling window backtesting.

Simulation is a way to model non-normal asset (and factor) distributions, and there are basically two types of simulation: historical simulation and Monte Carlo simulation. **Historical simulation** is relatively straightforward to perform: It uses past return data, and a random number generator picks observations from the historical series to simulate an asset's future returns. As such, historical simulation does suffer from the same issue as rolling window backtesting; both techniques assume that past asset returns provide sufficient guidance about future asset returns. Despite its limitations, historical simulation is widely used, particularly by banks for market risk analysis.

Monte Carlo simulation is more complex and computationally intensive compared with historical simulation, because each key decision variable in a Monte Carlo simulation requires an assumed statistical distribution, where a normal distribution is most frequently used as the default. However, as suggested by the previous discussions of non-normality, fat tails, and tail dependence, there is a clear need to incorporate non-normality in modeling. Researchers typically calibrate the parameters of the assumed distributional form (e.g., mean and standard deviation for a univariate normal distribution) using historical return data. The Monte Carlo simulation approach is popular because it is highly flexible and adaptable for solving high-dimensionality problems.

A properly designed simulation analysis is typically implemented by the following steps:

1. Determine the target variable that we want to understand. In investment research, the target variable is often the return on our investment strategy portfolio, $r_{p,t}$ (the return on portfolio p at time t), and we want to investigate its distribution.
2. Specify key drivers and decision variables that directly determine the value of our target variable. In an asset allocation strategy, key drivers/decision variables are the returns of each underlying asset, $r_{i,t}$ (the return on asset i at time t), in the overall portfolio and the weight, $\omega_{i,t}$ (the weight of asset i at time t), allocated to each asset. Once we know the returns and weights of all underlying assets, we can readily compute the return of our asset allocation strategy as $r_{p,t} = \sum_{i=1}^{K}\left(\varpi_{i,t} \times r_{i,t}\right)$.
3. Specify the number of trials (N) to run. Practitioners often choose a sufficiently large number of trials to get a useful distribution profile but not so many repetitions that the simulation exercise may consume too much computing time. In theory, exactly how to determine the appropriate number of iterations is a complex topic (for an example, see Ritter, Schoelles, Quigley, and Klein 2011).

In practice, researchers typically choose between 1,000 and 10,000 simulation runs, and the greater the number of trials, the more stable are the predictions of performance and variance of performance.

4 Define the distributional properties of the key drivers and decision variables in Step 2. This is the point where historical and Monte Carlo simulations diverge. In historical simulation, we assume that the distribution pattern of the historical data is sufficient to represent uncertainty in the future. Conversely, in Monte Carlo simulation, we must specify an exact functional form of the underlying statistical distribution for each key driver/decision variable. Note that researchers might specify different distributional functions (e.g., normal, lognormal, binomial) for different variables in a Monte Carlo simulation and thereby account for the impact of correlations and tail dependence in the multivariate distribution.

5 Use a random number generator to draw N random numbers—more specifically, **pseudo-random numbers**, or numbers that look random but are actually deterministic—for each key decision variable. A benefit of using a pseudo-random number generator is that once we fix the seed (the number that specifies the starting point for the pseudo-random number generator), we can re-run the simulation and verify the results; such reproducibility is an important property in scientific research. Note that different programming languages may use different random number generators. Even for the same language (with the same seed), the results may still be different with different versions and on different operating systems. However, the results should be qualitatively similar and thus, in this sense, are reproducible.

6 For each set of simulated drivers/decision variables, compute the value of the target variable. The value of the target variable is then saved for later analysis.

7 Repeat the same process from Steps 5 and 6 until completing the desired number of trials (N).

8 Now we have a set of N values of the target variable. In asset allocation simulations, this is the N likely returns of the investment strategy. The analyst can now calculate the typical performance measurements for the investment strategy, such as mean return, volatility, Sharpe ratio, and the various downside risk metrics. In simulation analysis, analysts typically focus on the downside risk profiles, so CVaR and maximum drawdown are appropriate.

Historical and Monte Carlo simulation techniques for evaluating an investment strategy will be demonstrated using the risk parity and benchmark strategies (comprising the eight underlying factor portfolios). To evaluate the results of these simulations, the performance of our investment strategies will be measured using the Sharpe ratio, as well as VaR, CVaR, and maximum drawdown metrics.

HISTORICAL SIMULATION 10

i. Explain Inputs and Decisions in Simulation, and Interpret a Simulation

In historical simulation, the key assumption is that past performance is a good indicator of future performance. So, to model randomness, the random number generator is used to randomly sample data from the historical return data in order to simulate future returns. First, a decision must be made about whether to sample from the historical

returns with replacement or without replacement. Random sampling with replacement, also known as **bootstrapping**, is often used in investment research, because the number of simulations needed is often larger than the size of historical dataset.

Before delving into the details of historical simulation, we first need to understand the difference between historical simulation and the rolling window backtesting technique demonstrated previously. Although both approaches rely on history to understand the future, they address the problem differently. Rolling window backtesting is deterministic. The investment manager constructs his or her investment strategy using historical data at each given point in time and then measures the strategy's performance in the next period. The same process is repeated consecutively from one past period to the next period. Thus, rolling window backtesting is designed to understand what the final outcome, in terms of performance, would have been if the investment manager had followed the specified trading rules (such as long/short hedge, Spearman rank IC, or some other trading algorithm).

Historical simulation (as well as Monte Carlo simulation) is non-deterministic and random (i.e., stochastic) in nature. Researchers randomly draw data from the historical track record—thus, not in a time-ordered sequence (which is another difference from rolling window backtesting). Hence, an important goal of simulation is to verify the investment performance obtained from backtesting. Moreover, simulation accounts for the randomness of the data in a different way from backtesting. As we will show, historical simulation randomly samples (with replacement) from the past record of asset returns, where each set of past monthly returns is equally likely to be selected. In contrast, to perform Monte Carlo simulation, researchers must first fit a multivariate joint probability distribution (e.g., normal or another type of distribution), where the past asset return data are used to calibrate the parameters. Once a particular model is fitted, we can randomly select data from the fitted distribution. Simulation is especially useful in measuring the downside risk of investment strategies, if the data are non-normal and tail dependence is captured properly.

Using the factor allocation strategies (BM and RP) for the eight factor portfolios as an example, a historical simulation can be designed in the following steps:

1. In this case, the target variables are the returns for the benchmark and the risk parity portfolios.
2. The key drivers/decision variables are the returns of the eight underlying factors. Note that for this simulation, the weights allocated to the eight factors are already known. For the BM portfolio, the weight is 1/8 for each factor. As noted earlier, the variance–covariance matrix for the eight factor portfolios is the key input for determining the weights assigned in the current period (May 2019) to each of the eight factor portfolios in the RP portfolio (see Panel A of Exhibit 17).
3. The simulation will be performed for N = 1,000 trials.
4. The historical simulation will be implemented using bootstrapped sampling. In this case, we will randomly draw a number from a uniform distribution (so there is equal probability of being selected) between 0 and 1.[2] Once a random number is generated, it can be mapped to a specific historical month. Note that we have a total of 374 months of historical factor return data (April 1988–May 2019). We can map a random number of a specific month by dividing the span of the uniform distribution by the number of months (1.0/374 = 0.00267). Therefore, if the random number is between 0 and 0.00267, then the first month

2 Technically, the random number generator will draw a random number that equals or is greater than 0 but is less than 1.

is selected. Similarly, if the random number generator picks a number between 0.00267 and 0.00535 (= 2 × 0.00267), then the second month is chosen, and so on.

5 The random number generator will then randomly draw 1,000 numbers from the uniform distribution between 0 and 1, and, as mentioned, sampling of the historical return data is with replacement. For example, as shown in Exhibit 28, the first five numbers generated are 0.59163, 0.32185, 0.76485, 0.89474, and 0.45431, which are then mapped to Months 222 (September 2006), 121 (April 1998), 287 (February 2012), 335 (February 2016), and 170 (May 2002), respectively. To be clear, months are mapped by dividing the random number by 0.00267, so Month 222 is determined as 0.59163/0.00267, Month 121 is 0.32185/0.00267, and so on.

Exhibit 28 The First Five Randomly Selected Months

Simulation #	Month	Random #	Month #	Earnings Yield	Book-to-Market	Earnings Growth	Momentum
1	9/30/2006	0.59163	222	2.5%	0.3%	(0.8%)	(0.0%)
2	4/30/1998	0.32185	121	0.1%	0.8%	(0.2%)	(0.5%)
3	2/29/2012	0.76485	287	(1.9%)	0.5%	1.7%	1.8%
4	2/29/2016	0.89474	335	2.5%	2.4%	(0.4%)	(1.5%)
5	5/31/2002	0.45431	170	6.3%	(3.3%)	1.8%	2.4%

Simulation #	Month	Random #	Month #	Earnings Revision	ROE	Debt/Equity	Earnings Quality
1	9/30/2006	0.59163	222	(0.8%)	2.5%	0.5%	(0.5%)
2	4/30/1998	0.32185	121	(0.1%)	(0.1%)	0.3%	1.6%
3	2/29/2012	0.76485	287	1.8%	(0.5%)	(2.1%)	(0.8%)
4	2/29/2016	0.89474	335	(1.5%)	1.2%	(1.2%)	1.3%
5	5/31/2002	0.45431	170	2.4%	6.4%	(0.7%)	(1.2%)

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

6 Once a given month is selected, the returns of the corresponding eight factor portfolios are used to represent one possible set of outcomes. Then, using the factor portfolio returns and the prespecified factor weights, we can compute the values of our target variables—the returns of the BM and RP portfolios. For example, the first trial picks the month of September 2006. The return of the benchmark portfolio is the equally weighted average of the eight factor returns, or 0.46% (= 0.125 × 2.5% + 0.125 × 0.3% + 0.125 × –0.8% + 0.125 × 0.0% + 0.125 × –0.8% + 0.125 × 2.5% + 0.125 × 0.5% + 0.125 × –0.5%).

To compute the return on the risk parity portfolio, we need the weights allocated to each of the eight factors for the current month (May 2019). As shown in Exhibit 29, for the first trial, September 2006, the weighted average return of the risk parity portfolio is 0.2% (actually, 0.17%). It should be clear that each trial in the historical simulation assumes the simulated returns of the eight factors follow the same patterns observed in the sampled month—in this case, September 2006.

Exhibit 29 How to Compute the Return of the Risk Parity Portfolio, Historical Simulation #1

Asset (Factor)	September 2006 Return	May 2019 Weight	Weighted Return
Earnings Yield	2.5%	6.0%	0.2%
Book-to-Market	0.3%	30.3%	0.1%
Earnings Growth	(0.8%)	11.7%	(0.1%)
Momentum	(0.0%)	5.2%	(0.0%)
Earnings Revision	(0.8%)	10.4%	(0.1%)
ROE	2.5%	6.3%	0.2%
Debt/Equity	0.5%	9.6%	0.0%
Earnings Quality	(0.5%)	20.4%	(0.1%)
Risk Parity Portfolio			**0.2%**

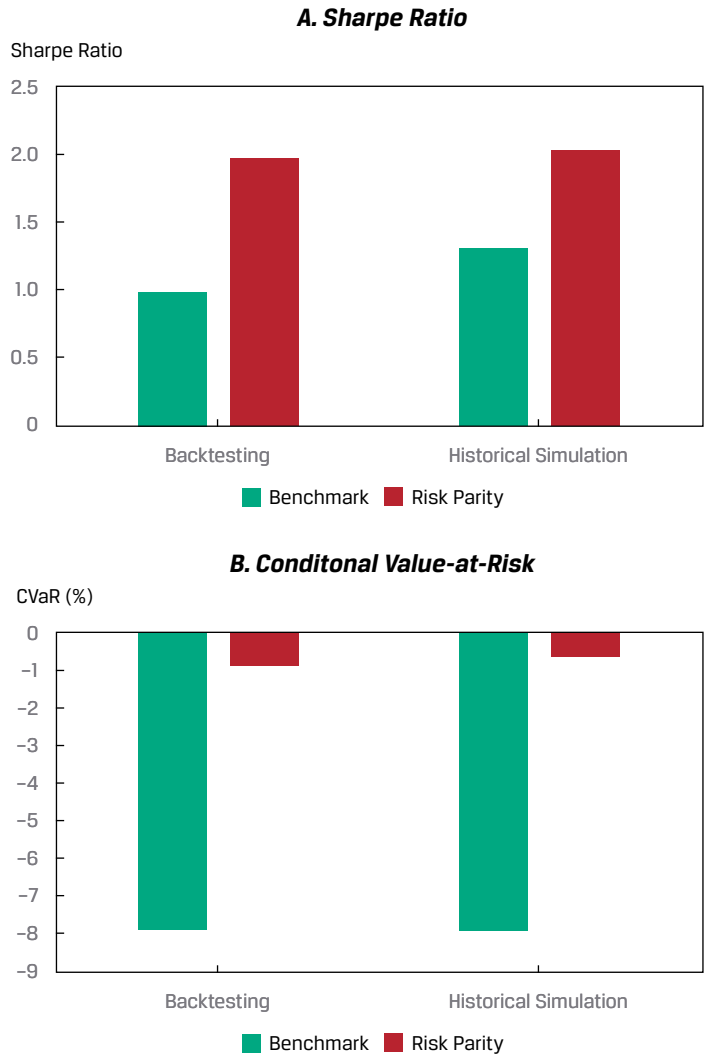
Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES.

7 The same simulation process (from Steps 5 to 6) is repeated for 1,000 trials, and then we have a collection of 1,000 simulated returns for the benchmark and risk parity portfolios.

8 Finally, equipped with these 1,000 return scenarios, we can calculate the performance metrics of interest (Sharpe ratio, CVaR, etc.) and plot the distributions of the benchmark and risk parity portfolio returns.

As shown in Panel A of Exhibit 30, the results of the historical simulation (over the 1,000 iterations) suggest that the Sharpe ratios of the BM and RP strategies are largely in line with the rolling window backtesting method demonstrated previously. In particular, the RP portfolio outperforms the naive BM portfolio in terms of Sharpe ratio using both methodologies. Similarly, as shown in Panel B, both methodologies indicate that the RP portfolio carries substantially less downside risk, measured by CVaR, than the BM portfolio carries.

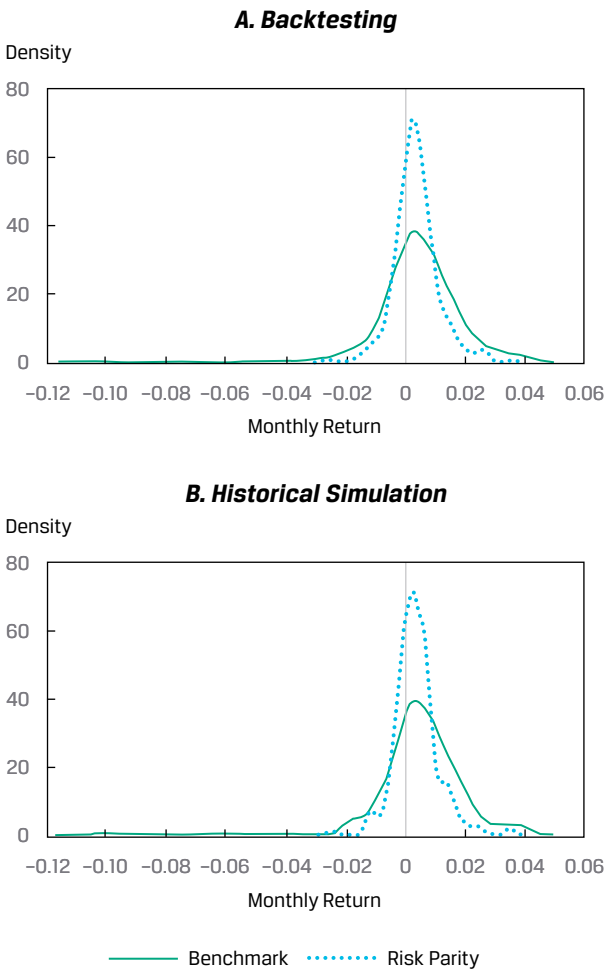
Exhibit 30 Comparing Historical Simulation with Backtesting



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

In addition to capturing downside risk with a single number (e.g., CVaR), we can also plot the estimated probability distribution of returns for our two investment strategies. Panel A of Exhibit 31, plots the estimated probability distribution of returns for the BM and RP portfolios using backtested returns, whereas Panel B shows the estimated return distribution plots using the historical simulated returns. We can observe a broadly similar pattern between them. Both the backtesting and historical simulation approaches suggest that the RP portfolio returns are less volatile and more skewed to the right with lower downside risk (i.e., lower standard deviation and thinner tails) than the BM portfolio returns.

Exhibit 31 Estimated Distribution Plots: Backtesting and Historical Simulation



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

However, an important issue with this historical simulation (and historical simulation generally) is that the path dependency of the time series of factor returns (and asset returns, generally) is not preserved. This deficiency can be addressed with the Monte Carlo simulation method.

11 MONTE CARLO SIMULATION

Monte Carlo simulation follows similar steps as historical simulation but with a few key differences.

First, in historical simulation, each random variable of interest (key driver or decision variable) is randomly drawn from historical data (Step 4 in the prior discussion of historical simulation). In a Monte Carlo simulation, we need to specify a functional form for each decision variable. The exploratory data analysis presented earlier—focusing on moments (i.e., mean, standard deviation, skewness, kurtosis) and tail dependence—can help us understand the empirical distribution of key drivers and

decision variables. The usefulness of the Monte Carlo simulation technique critically depends on whether the functional form of the statistical distribution that we specify accurately reflects the true distribution of the underlying data. Because the true distribution of the data is unknown, we need to be aware of the fact that our model, like all models, only provides guidance and will not be perfect.

Second, once we specify the functional form(s) (note that different key drivers/decision variables may have different functional forms), regression and optimization techniques are used to estimate the parameters (i.e., mean, standard deviation, skewness, kurtosis) underlying the statistical distribution. This step is typically called model calibration. Although it may sound difficult, computer programming languages widely used by data scientists (e.g., R, Python, Matlab) can readily fit empirical data to a multivariate normal distribution with a few lines of code. For example, the fMultivar package in the R programming language offers many useful functions to fit data and simulate from multivariate distributions.

When we choose the functional form of the statistical distribution, we need to account for the following considerations:

- The distribution should be able to reasonably describe the key empirical patterns of the underlying data. For example, asset returns typically follow a bell curve pattern (e.g., the value factor returns observed in Exhibit 19); therefore, the normal distribution and Student's *t*-distribution are often used as a first-cut approximation.
- It is equally critical to account for the correlations between multiple key drivers and decision variables. In the case of asset or factor allocation strategies, as shown previously, the returns from multiple factors are clearly correlated; therefore, we need to specify a multivariate distribution rather than modeling each factor or asset on a standalone basis.
- The complexity of the functional form and number of parameters that determine the functional form are equally important. This is the trade-off between model specification error and estimation error. We can specify a highly complex model with many parameters (all of which need to be estimated/calibrated from historical data) that describe the empirical properties of the data well. However, given limited historical data, we may not be able to estimate all the underlying parameters with sufficient precision. Such models tend to have low specification errors, but they suffer from large estimation errors. On the other extreme, overly simplistic models require fewer parameters (therefore, they might have low estimation errors), but they may not fit the data well (because they are misspecified). You should recognize this phenomenon as the bias–variance trade-off, introduced in earlier readings on machine learning and big data projects.

For asset or factor allocation strategy simulation, the distribution of asset or factor returns is typically modeled as a multivariate normal distribution—as a first-cut approximation—which captures some of the key properties of the underlying data reasonably well. More importantly, a multivariate normal distribution can be fully specified with only a few key parameters—the mean, the standard deviation, and the covariance matrix. For K assets, we need to estimate K mean returns, K standard deviations, and $[K \times (K - 1)]/2$ correlations.

However, we have to be aware that the multivariate normal distribution does not fully account for the empirical characteristics of (negative) skewness, excess kurtosis, and tail dependence in the data. We will address these non-normal distribution properties shortly, when we cover sensitivity analysis.

Continuing with the same benchmark and risk parity allocation strategies, we now perform the Monte Carlo simulation using the following steps:

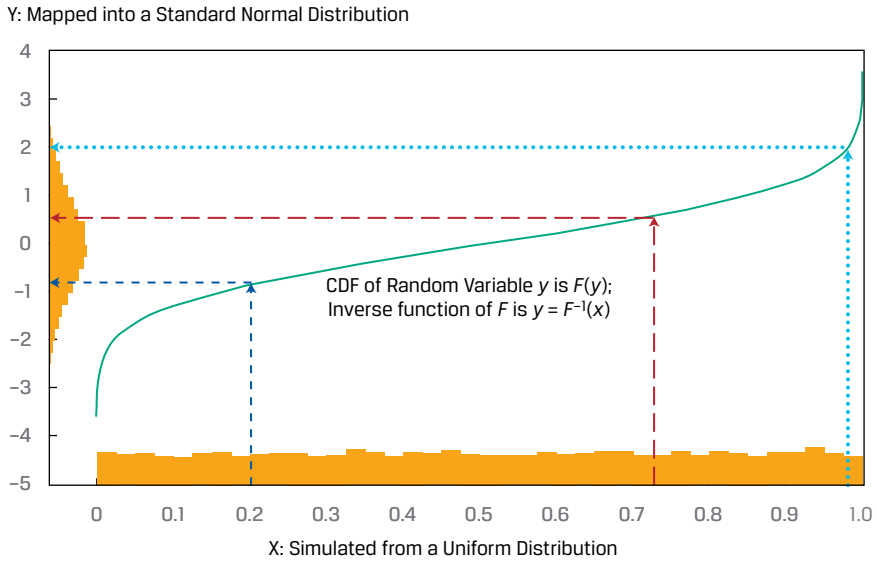
1. The target variables are the returns of the benchmark and the risk parity portfolios.
2. The key drivers/decision variables are the returns of the eight underlying factor portfolios.
3. We will perform the simulation using 1,000 trials (the same as for the historical simulation exercise).
4. We choose the multivariate normal distribution as our initial functional form. We calibrate the model—calculate the eight factor portfolio mean returns, the eight standard deviations, and the 28 elements of the covariance matrix—using the 374 months of historical factor return data (April 1988–May 2019).
5. The calibrated multivariate normal distribution is then used to simulate the future factor returns. The process by which this occurs will be described (roughly) first in the simpler context of one normally distributed random variable and then in the current context of a multivariate normal distribution of eight random variables.

   In the case of one normally distributed (i.e., univariate normal) random variable, the simulation process for future returns can be thought of as follows: First, the random number generator selects a number from the uniform distribution between 0 and 1 (think of this random number as an x-axis coordinate). Next, that number is directly mapped onto the random variable's cumulative probability distribution function (cdf), which also ranges from 0 to 1. Finally, the y coordinate of the point on the cdf curve is used to determine the value of the random variable for that trial. The process of converting a randomly generated uniformly distributed number into a simulated value of a random variable of a desired distribution is known as the **inverse transformation method**. It is depicted in Exhibit 32, where randomly generated values from the uniform distribution (x's) are mapped into a standard normal distribution to determine simulated values (y's) of the random variable.

   It is a remarkable fact that random observations from any distribution can be produced using the uniform random variable with endpoints 0 and 1. To see why this is so, consider the inverse transformation method that converts from the uniform distribution to the standard normal distribution (as in Exhibit 32). Suppose we are interested in obtaining random observations for a random variable y, with cumulative distribution function $F(y)$. Recall that $F(y)$ evaluated at y is a number between 0 and 1. For instance, to produce a random outcome of 2.00 from the standard normal distribution (i.e., a +2 × sigma event), the uniform random number generator must select the number 0.98 (chosen from the range of 0 and 1), since $F(2.00) = 0.98$. Define the inverse function of F—call it $y = F^{-1}(x)$—that transforms the input number 0.98 into the random outcome of 2.00; so, $F^{-1}(0.98) = 2.00$. To generate random observations on variable y, the steps are to generate a uniform random number, x, between 0 and 1 using the random number generator and then to evaluate $F^{-1}(x)$ to obtain a random observation of y.

   Random observation generation is a field of study in itself, and we have briefly discussed the inverse transformation method here just to illustrate a point. As a generalist, you do not need to address the technical details of converting random numbers into random observations, but you do need to know that random observations from any distribution can be generated using a uniform random variable.

Exhibit 32 Inverse Transformation: From Uniform Distribution to Standard Normal Distribution



Sources: Wolfe Research Luo's QES.

In the context of a multivariate normal distribution with eight random variables, the process is considerably more complex to visualize and explain. Suffice it to say, in this case, eight randomly generated numbers from the uniform distribution are mapped onto a point on the joint cumulative probability distribution function (actually, a high-dimensional cdf space), and this point is used to jointly determine the values of the eight factor returns in this trial.

Exhibit 33 shows the first five simulated sets of returns for the eight underlying factors that result from implementing the inverse transformation method.

Exhibit 33 Monte Carlo Simulation: First Five Simulated Months of Factor Returns Using a Multivariate Normal Distribution

Simulation #	Earnings Yield	Book-to-Market	Earnings Growth	Momentum	Earnings Revision	ROE	Debt/Equity	Earnings Quality
1	(3.2%)	(3.1%)	(0.2%)	0.7%	2.3%	(3.3%)	(1.7%)	1.9%
2	(0.0%)	3.5%	0.9%	(0.4%)	0.9%	(2.4%)	(3.5%)	(0.2%)
3	0.7%	(1.8%)	2.9%	3.8%	2.5%	1.3%	(0.8%)	(0.0%)
4	9.7%	(0.5%)	1.2%	3.8%	(0.9%)	7.6%	(3.7%)	1.6%
5	1.7%	0.2%	2.9%	(0.2%)	3.0%	0.2%	(0.9%)	0.2%

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

6 Once the returns of the eight factor portfolios are simulated, we can compute the values of our target variables—the returns of the two factor allocation portfolios. For example, for the first simulated set of returns, the benchmark

portfolio (with equally weighted factor returns) delivers a monthly return of –0.83% (= 0.125 × –3.2% + 0.125 × –3.1% + 0.125 × –0.2% + 0.125 × 0.7% + 0.125 × 2.3% + 0.125 × –3.3% + 0.125 × –1.7% + 0.125 × 1.9%).

Similarly, using the RP allocation factor weights for the current month, May 2019, shown in Exhibit 29, the risk parity portfolio return for this simulated month is –0.86% (= 0.06 × –3.2% + 0.303 × –3.1% + 0.117 × –0.2% + 0.052 × 0.7% + 0.104 × 2.3% + 0.063 × –3.3% + 0.096 × –1.7% + 0.204 × 1.9%).

7 Next, we repeat Steps 5 and 6 for 1,000 trials to get a collection of 1,000 returns for the benchmark and risk parity portfolios.

8 Finally, we assess the performance and risk profiles of our two investment strategies from the 1,000 simulated returns.

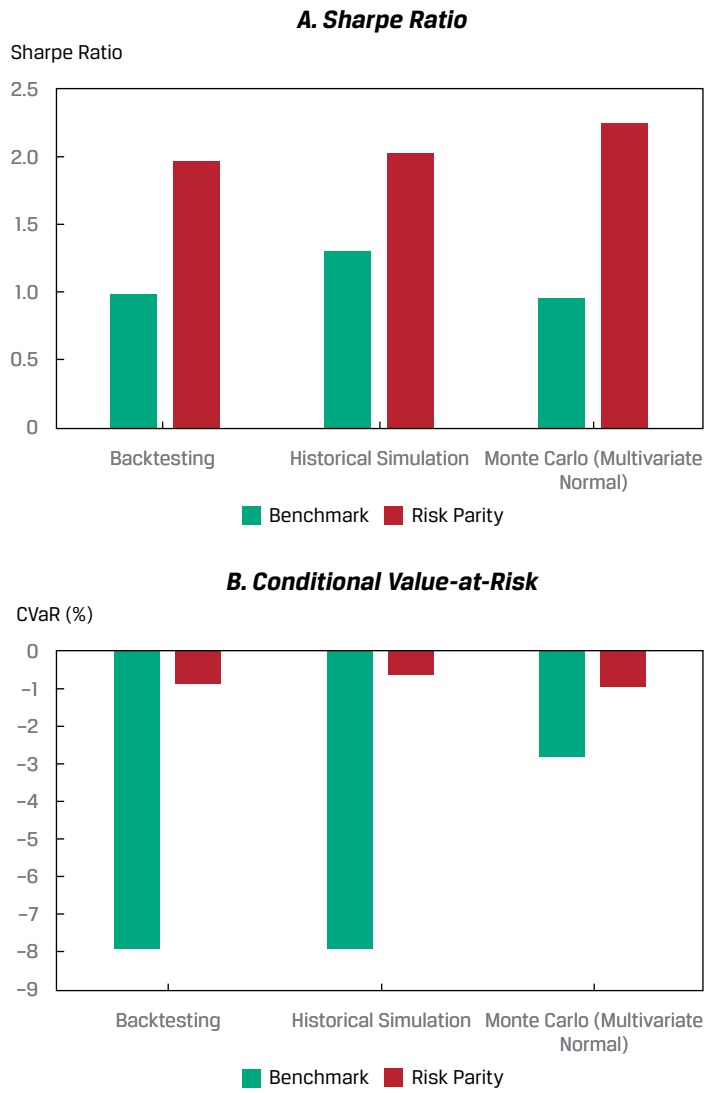
EXAMPLE 9

How to Interpret the Results from Historical and Monte Carlo Simulations

Exhibit 34 shows the Sharpe ratios (Panel A) and downside risk measures (CVaRs, Panel B) for the returns of the benchmark and risk parity allocation strategies based on rolling window backtesting, historical simulation, and Monte Carlo simulation of the returns on the eight underlying factor portfolios.

Discuss the similarities and differences among the three approaches for simulated performance of the benchmark and risk parity portfolios.

Exhibit 34 Comparing Backtesting, Historical Simulation, and Monte Carlo Simulation for the BM and RP Strategies



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

Solution:

Note the backtesting approach provides realistic performance metrics assuming investors have been following the same trading rules throughout the past periods under investigation. The two simulation analyses are complementary to backtesting and deliver additional insights. In particular, they account for the random nature of investment data in different ways. Historical simulation randomly samples (with replacement) from the past record of asset returns, where each set of past monthly returns is equally likely to be selected. Monte Carlo simulation randomly samples from an assumed multivariate joint probability distribution (e.g., normal or another type of distribution), where the past record of asset returns is used to calibrate the parameters of the multivariate distribution. Therefore, these simulation methods are used to independently verify the results from the rolling window backtesting.

As shown in Panel A of Exhibit 34, the Sharpe ratio appears relatively insensitive to the simulation and backtesting methods used, with the RP strategy outperforming the BM strategy by nearly the same margin for each method. In contrast, CVaR seems to be sensitive to how randomness is treated. In particular, the Monte Carlo simulation appears to understate the downside risk of the BM strategy compared with both rolling window backtesting and historical simulation methods (Panel B). Since the factor returns are negatively skewed with fat tails (i.e., excess kurtosis), the multivariate normal distribution assumption is likely to be underestimating the true downside risk of the BM strategy. This underestimation of risk appears only for the BM strategy because factor risks and correlations are not properly accounted for in the naive (equal) weighting scheme. Conversely, in this case, the risk parity strategy is robust to a non-normal factor return distribution, resulting in a portfolio with considerably lower downside risk.

12 SENSITIVITY ANALYSIS

j Demonstrate the Use of Sensitivity Analysis

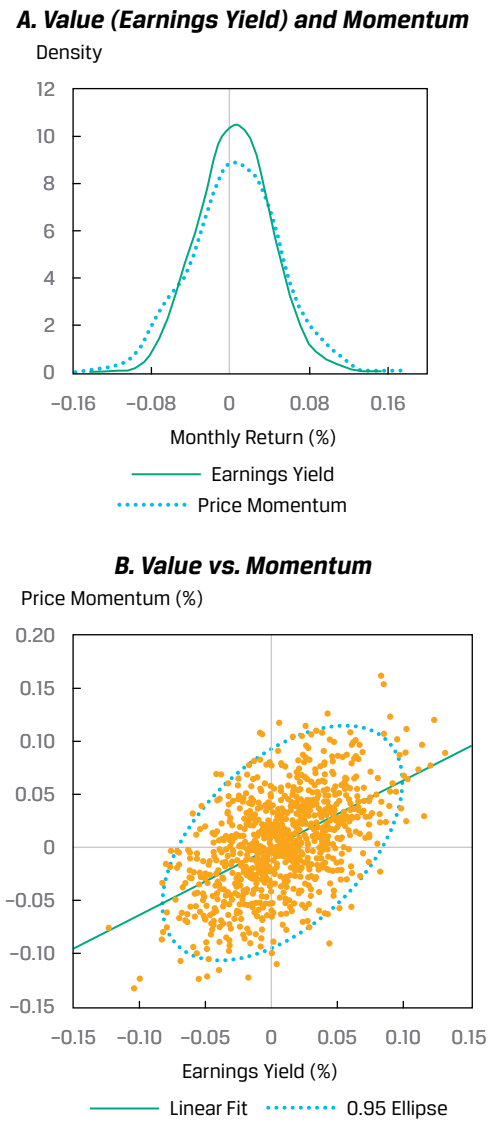
In addition to simulation, **sensitivity analysis**—a technique for exploring how a target variable (e.g., portfolio returns) and risk profiles are affected by changes in input variables (e.g., the distribution of asset or factor returns)—can be implemented to help managers further understand the potential risks and returns of their investment strategies.

The Monte Carlo simulation just described fits a multivariate normal distribution to the factor returns, which is a sensible first approximation. On the positive side, this preserves the cross-sectional integrity of the factor returns. Compared with other, more complex distribution functions, the multivariate normal distribution requires fewer parameters to be estimated from historical data. However, the value of the simulation results depends crucially on whether the multivariate normal distribution is the correct functional form or whether it is at least a reasonable proxy for the true distribution. Unfortunately, in investment management, the true functional form of our variables of interest (here, the factor returns) is almost never known. This is the main reason why one needs to perform a sensitivity analysis.

Despite the simplicity and wide adoption in practice, the multivariate normal distribution assumption fails to account for the various empirical properties we observed for factor returns, including negative skewness and fat tails (see Exhibit 20). For example, Panel A of Exhibit 35 shows plots of the distribution of the earnings yield and price momentum factor returns derived by Monte Carlo simulation from the multivariate normal distribution model fitted in the previous exercise. Clearly, both factors appear to follow a normal distribution quite well. However, although the average returns and the volatilities of the two simulated factors are similar to the actual observed returns and volatilities, the negative skewness and fat tails observed empirically (see Panel A of Exhibit 19) now completely disappear.

Furthermore, the correlation between the two factors (see Panel B of Exhibit 35) also appears to be nicely behaved. A 95% ellipse circle contains the vast majority of simulated return data, which is somewhat different from the actual observed returns in Panel B of Exhibit 19.

Exhibit 35 Distribution of Selected Monte Carlo Simulated Factor Returns (Assuming Multivariate Normal Distribution)



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

As a robustness test, instead of the multivariate normal distribution used previously, we conduct a sensitivity analysis. We do this by fitting our factor return data to a different distribution, a multivariate skewed Student's *t*-distribution, and then we repeat the Monte Carlo simulation accordingly. The Student's *t*-distribution is also symmetric (similar to a normal distribution), but it has fatter tails than a normal distribution. To capture the negative skewness in the factor returns, we use the skewed generalized *t*-distribution; it is a natural extension of the multivariate normal distribution but has the ability to account for the skewness and the excess kurtosis often observed in factor and asset return data.

The multivariate skewed *t*-distribution is mathematically more complex and requires estimating a larger number of parameters than a normal distribution. Therefore, although the assumption of this distribution may better approximate the statistical properties of asset return data, there is no guarantee that it will deliver more accurate

predictions than the traditional multivariate normal assumption. Again, it is fairly straightforward to fit our empirical data (i.e., 374 months of factor returns) into a multivariate skewed *t*-distribution using any of the standard data science programming language packages.

With the goal of determining the sensitivity of our target variables (the returns of the benchmark and the risk parity portfolios) to the new factor return distribution assumption, the procedure for the new Monte Carlo simulation is almost identical to the one performed previously. The only two exceptions are Steps 4 and 5. In Step 4, instead of fitting the data to a multivariate normal distribution, we calibrate our model to a multivariate skewed *t*-distribution. In Step 5, we simulate 1,000 sets of factor returns from this new distribution function. Then, as before, we can assess the performance and risk profiles of our investment strategies from the 1,000 new simulated returns.

Exhibit 36 shows the first five sets of simulated factor returns from this new model. As previously, we compute the values of our target variables for each set of simulated factor returns and then assess their performance and risk characteristics. For the first set of factor returns, the equal-weighted (i.e., 0.125 for each factor) benchmark portfolio achieves a monthly return of 1.21%, and the equal-risk-weighted (i.e., factor weights for May 2019 in Exhibit 29) risk parity portfolio delivers a return of 0.75%.

Exhibit 36 First Five Simulated Months of Factor Returns Using Multivariate Skewed t-Distribution

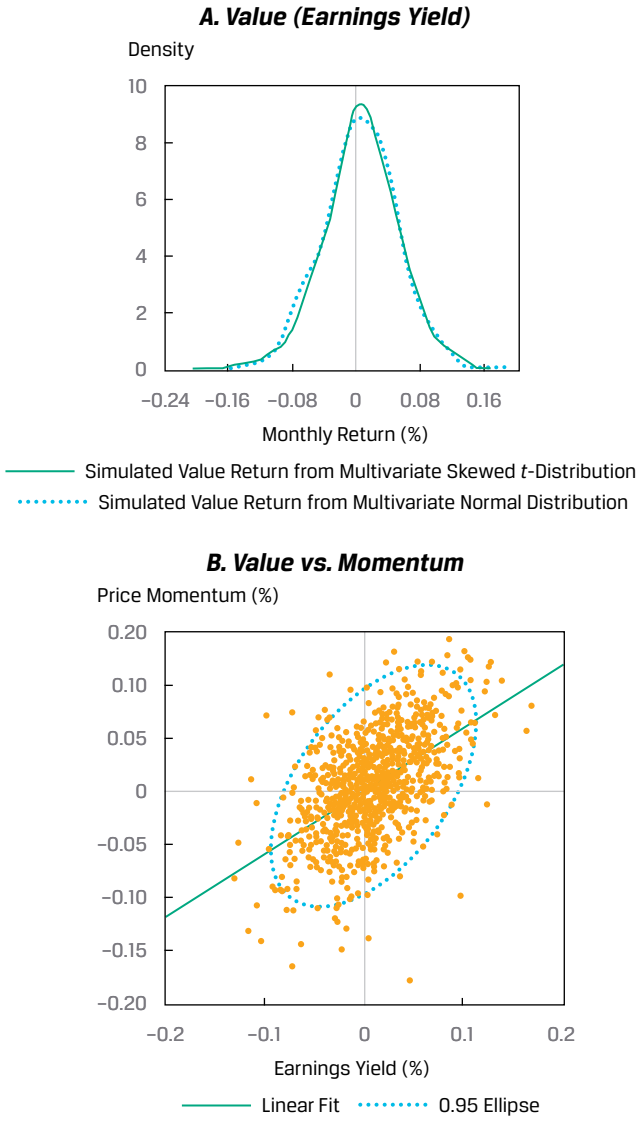
Simulation #	Earnings Yield	Book-to-Market	Earnings Growth	Momentum	Earnings Revision	ROE	Debt/Equity	Earnings Quality
1	2.0%	0.3%	1.7%	3.1%	2.0%	0.9%	0.2%	(0.5%)
2	1.8%	(1.4%)	0.2%	4.9%	1.8%	2.7%	0.4%	(0.1%)
3	(0.6%)	0.2%	(1.0%)	(0.1%)	0.4%	1.5%	1.6%	0.9%
4	11.2%	2.6%	1.8%	1.5%	2.2%	9.6%	(2.9%)	(1.9%)
5	(3.9%)	(1.3%)	0.9%	0.9%	0.8%	(3.5%)	2.9%	0.2%

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

In Panel A of Exhibit 37, we compare the distributions of the simulated earnings yield factor returns from the fitted multivariate normal and multivariate skewed t-distribution models, respectively. The skewed *t*-model, which has five parameters,[3] shows more negative skewness with potentially more left-tail surprises compared with the other model. Similarly, Panel B illustrates the pairwise scatterplot for earnings yield and price momentum factors simulated from the new multivariate skewed *t*-distribution model. The new plot appears to have more outliers outside of the 95% ellipse (the ellipse is based on the multivariate normal distribution), indicating that the skewed *t*-model captures fat tails better than a normal distribution does (see Panel B of Exhibit 35).

3 The five parameters in specifying a skewed *t*-distribution are scale, location, skewness, and two parameters controlling for kurtosis.

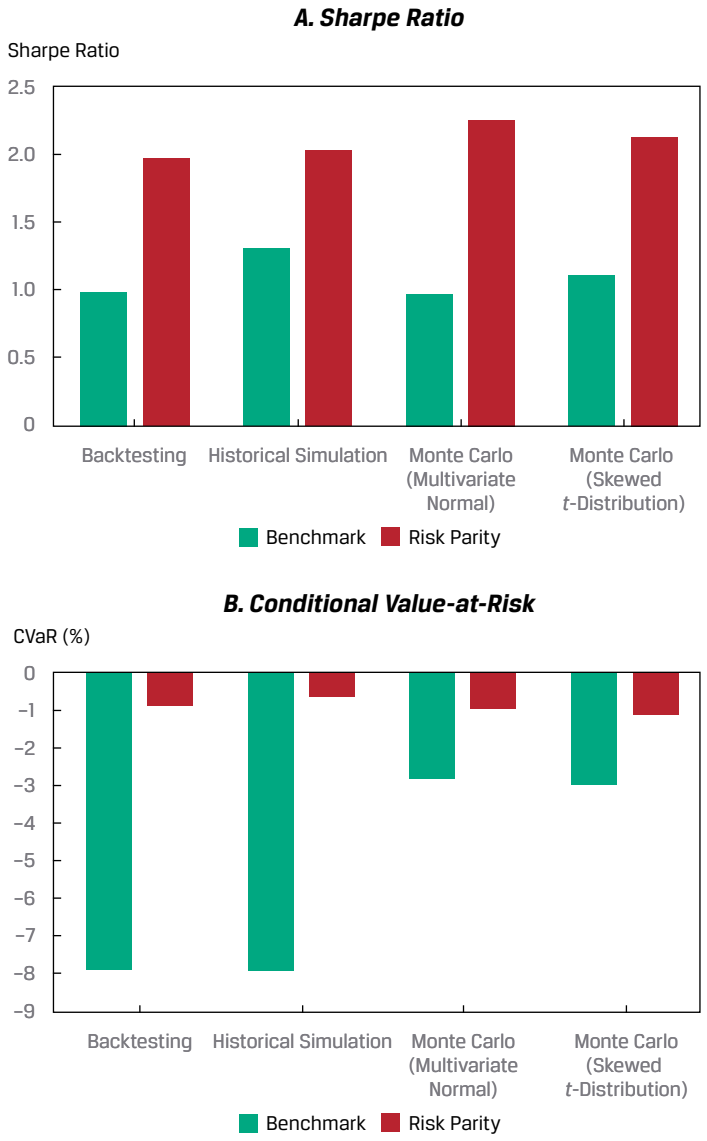
Exhibit 37 Distribution of Select Simulated Factor Returns (Multivariate Skewed *t*-Distribution)



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

Turning to the performance and risk profiles of our investment strategies, as shown in Panel A of Exhibit 38, the Sharpe ratio appears insensitive to any of the particular simulation methods used: They all consistently suggest that the risk parity factor allocation strategy outperforms the benchmark strategy. Downside risk (expressed as CVaR), however, appears quite sensitive to the choice of simulation approach for the BM strategy but not very sensitive for the RP strategy (Panel B). Focusing on the BM strategy, the CVaR results from historical simulation and rolling window backtesting resemble each other very closely. The CVaR results of the skewed *t*-distribution and multivariate normal Monte Carlo simulations are also very similar: They both underestimate the downside risk of the BM strategy. This finding suggests that additional sensitivity analyses should be run with different functional forms for the factor return distribution.

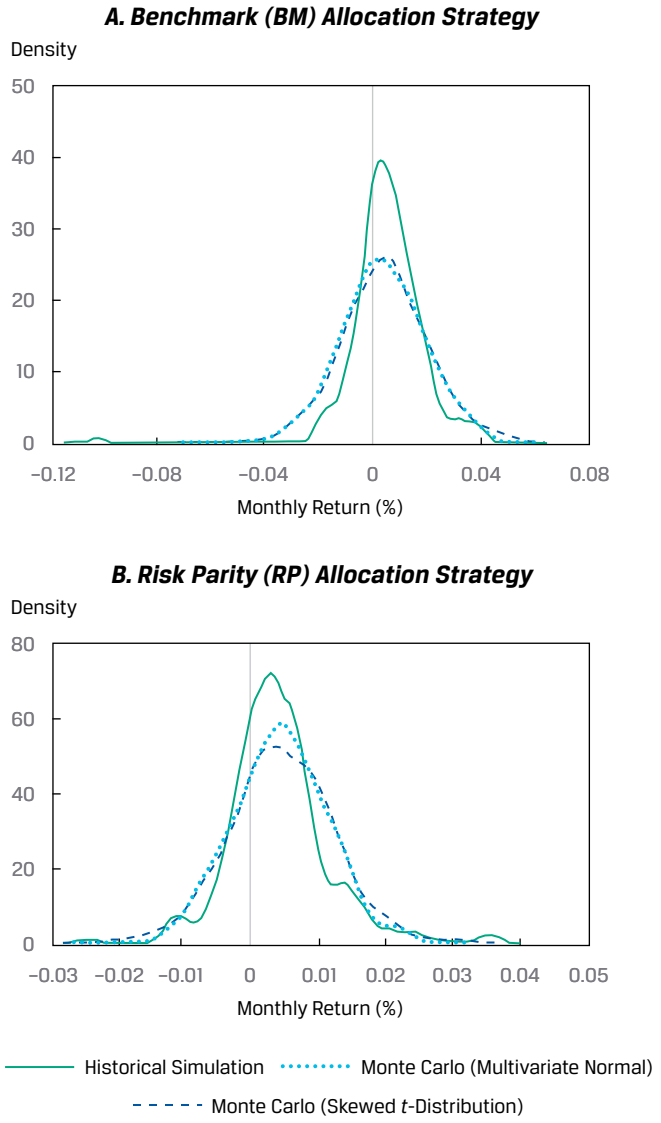
Exhibit 38 Comparing Simulation Methods with Backtesting



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

Using estimated probability density plots, shown in Panel A of Exhibit 39, it can be seen that the difference between the historical simulation and the two Monte Carlo methods is rather large for the BM strategy. Given the negative skewness and excess kurtosis of the BM strategy's returns, which is apparent from the shape of the historical simulation return distribution, it is not surprising that the two Monte Carlo simulations fail to account for this left-tail risk property. Conversely, since the distribution of the RP strategy's returns is relatively symmetric and without much excess kurtosis, all three simulation methods provide a fairly similar picture (Panel B).

Exhibit 39 Estimated Distribution Plots for BM and RP Strategies Using Three Different Simulations



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

EXAMPLE 10

Simulating the Performance of Factor Allocation Strategies

Earlier, Sarah Koh presented her team's backtesting results for the factor-based allocation strategies being considered by an important client, SWF Fund. Now, while presenting the simulation results for these same strategies, SWF Fund's investment committee asks Koh the following questions:

1 The following are caveats regarding the use of rolling window backtesting in assessing investment strategies *except*:

A this technique implicitly assumes that the same pattern of past performance is likely to repeat itself over time.

B this technique may not fully account for the dynamic nature of financial markets and potentially extreme downside risks.

C this technique is intuitive, because it mimics how investing is done in reality—that is, forming ideas, testing strategies, and implementing periodically.

2 Which one of the following statements is *false*?

A Volatility clustering is when a period of high volatility is likely to be followed by another period of high volatility or when a low-volatility period is likely to be followed by another low-volatility period.

B The tail dependence coefficient is a type of correlation coefficient that focuses on co-movements in the tails of two random variables.

C If the distribution of factor returns is non-normal, then just mean and standard deviation are needed to capture the randomness in the data.

3 Which of the following situations is *most likely* to involve data snooping?

A A researcher specifies an acceptable proportion (q-value) of significant results that can be false positives and establishes, on the basis of the q-value and ranked p-values, a critical value for reported p-values. She will only accept as significant tests with p-values below the critical value.

B A researcher tries many different modeling techniques, backtesting each of them, and then picking the best performing model without accounting for model selection bias.

C A researcher sets a relatively high hurdle, a t-statistic greater than 3.0×, for assessing whether a newly discovered factor is statistically significant.

4 Which of the following situations is *least likely* to involve scenario analysis?

A Simulating the performance and risk of investment strategies by first using stocks in the Nikkei 225 index and then using stocks in the TOPIX 1000 index.

B Simulating the performance and risk of investment strategies in both "trade agreement" and "no-trade-agreement" environments.

C Simulating the performance and risk of investment strategies in both high-volatility and low-volatility environments.

5 Which one of the following statements concerning historical simulation and Monte Carlo simulation is *false*?

A Historical simulation randomly samples (with replacement) from the past record of asset returns, where each set of past monthly returns is equally likely to be selected.

B Neither historical simulation nor Monte Carlo simulation makes use of a random number generator.

C Monte Carlo simulation randomly samples from an assumed multivariate joint probability distribution where the past record of asset returns is used to calibrate the parameters of the multivariate distribution.

6 Which one of the following statements concerning Monte Carlo simulation is *false*?

A When simulating multiple assets (factors) whose returns are correlated, it is crucial to specify a multivariate distribution rather than modeling each asset on a standalone basis.

B The inverse transformation method is a process for converting a randomly generated number into a simulated value of a random variable.

C The Monte Carlo simulation process is deterministic and non-random in nature.

7 Which of the following situations concerning simulation of a multifactor asset allocation strategy is *most likely* to involve sensitivity analysis?

A Changing the specified multivariate distribution assumption from a normal to a skewed t-distribution to better account for skewness and fat tails.

B Splitting the rolling window between periods of recession and non-recession.

C Splitting the rolling window between periods of high volatility and low volatility.

Solution to 1:

C is correct, since it is not a caveat in using rolling window backtesting. A and B are incorrect because they are caveats in the use of this technique.

Solution to 2:

C is correct, since statement C is false. A and B are incorrect because those statements are true.

Solution to 3:

B is correct, since this situation most likely involves data snooping. A and C are incorrect because these are approaches to avoiding data snooping.

Solution to 4:

A is correct, since there is no structural break or different structural regime. B and C are incorrect because they involve structural breaks/different structural regimes and thus represent different scenarios.

Solution to 5:

B is correct, since this statement is false. A and C are incorrect because they are true statements about historical and Monte Carlo simulation, respectively.

Solution to 6:

C is correct, since this statement is false. A and B are incorrect because they are true statements about Monte Carlo simulation.

Solution to 7:

A is correct, since this choice represents sensitivity analysis. B and C are incorrect because these choices represent scenario analysis.

SUMMARY

In this reading, we have discussed on how to perform rolling window backtesting—a widely used technique in the investment industry. Next, we described how to use scenario analysis and simulation along with sensitivity analysis to supplement backtesting, so investors can better account for the randomness in data that may not be fully captured by backtesting.

- The main objective of backtesting is to understand the risk–return trade-off of an investment strategy, by approximating the real-life investment process.
- The basic steps in a rolling window backtesting include specifying the investment hypothesis and goals, determining the rules and processes behind an investment strategy, forming an investment portfolio according to the rules, rebalancing the portfolio periodically, and computing the performance and risk profiles of the strategy.
- In the rolling window backtesting methodology, researchers use a rolling window (or walk-forward) framework, fit/calibrate factors or trade signals based on the rolling window, rebalance the portfolio periodically, and then track the performance over time. Thus, rolling window backtesting is a proxy for actual investing.
- There are two commonly used approaches in backtesting—long/short hedged portfolio and Spearman rank IC. The two approaches often give similar results, but results can be quite different at times. Choosing the right approach depends on the model building and portfolio construction process.
- In assessing backtesting results, in addition to traditional performance measurements (e.g., Sharpe ratio, maximum drawdown), analysts need to take into account data coverage, return distribution, factor efficacy, factor turnover, and decay.
- There are several behavioral issues in backtesting to which analysts need to pay particular attention, including survivorship bias and look-ahead bias.
- Risk parity is a popular portfolio construction technique that takes into account the volatility of each factor (or asset) and the correlations of returns between all factors (or assets) to be combined in the portfolio. The objective is for each factor (or asset) to make an equal (hence "parity") risk contribution to the overall or targeted risk of the portfolio.
- Asset (and factor) returns are often negatively skewed and exhibit excess kurtosis (fat tails) and tail dependence compared with normal distribution. As a result, standard rolling window backtesting may not be able to fully account for the randomness in asset returns, particularly on downside risk.
- Financial data often face structural breaks. Scenario analysis can help investors understand the performance of an investment strategy in different structural regimes.
- Historical simulation is relatively straightforward to perform but shares pros and cons similar to those of rolling window backtesting. For example, a key assumption these methods share is that the distribution pattern from the historical data is sufficient to represent the uncertainty in the future. Bootstrapping (or random draws with replacement) is often used in historical simulation.
- Monte Carlo simulation is a more sophisticated technique than historical simulation is. In Monte Carlo simulation, the most important decision is the choice of functional form of the statistical distribution of decision variables/return

drivers. Multivariate normal distribution is often used in investment research, owing to its simplicity. However, a multivariate normal distribution cannot account for negative skewness and fat tails observed in factor and asset returns.

- The Monte Carlo simulation technique makes use of the inverse transformation method—the process of converting a randomly generated uniformly distributed number into a simulated value of a random variable of a desired distribution.
- Sensitivity analysis, a technique for exploring how a target variable and risk profiles are affected by changes in input variables, can further help investors understand the limitations of conventional Monte Carlo simulation (which typically assumes a multivariate normal distribution as a starting point). A multivariate skewed *t*-distribution takes into account skewness and kurtosis but requires estimation of more parameters and thus is more likely to suffer from larger estimation errors.

REFERENCES

DeMiguel, V., L. Garlappi, and R. Uppal. 2007. "*Optimal Versus Naïve Diversification: How Inefficient Is the 1/N Portfolio Strategy?*" London Business School Working Paper. http://faculty.london.edu/avmiguel/DeMiguel-Garlappi-Uppal-RFS.pdf.

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

Fama, E., and J. D. MacBeth. 1973. "Risk, Return, and Equilibrium: Empirical Tests." *Journal of Political Economy* 81 (3): 607–36.

Jegadeesh, N., and S. Titman. 1993. "Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency." *Journal of Finance* 48 (1): 65–91.

Ritter, F. E., M. J. Schoelles, K. S. Quigley, and L. C. Klein. 2011. "Determining the Number of Simulation Runs: Treating Simulations as Theories by Not Sampling Their Behavior." In *Human-in-the-Loop Simulations: Methods and Practice*, ed. Rothrock, L., and S. Narayanan.97–116. Berlin: Springer.

PRACTICE PROBLEMS

The following information relates to Questions 1–8

Emily Yuen is a senior analyst for a consulting firm that specializes in assessing equity strategies using backtesting and simulation techniques. She is working with an assistant, Cameron Ruckey, to develop multifactor portfolio strategies based on nine factors common to the growth style of investing. To do so, Yuen and Ruckey plan to construct nine separate factor portfolios and then use them to create factor-weighted allocation portfolios.

Yuen tasks Ruckey with specifying the investment universe and determining the availability of appropriate reporting data in vendor databases. Ruckey selects a vendor database that does not provide point-in-time data, then makes adjustments to include point-in-time constituent stocks and assumes a reporting lag of four months.

Next, Yuen and Ruckey run initial backtests by creating a stock portfolio and calculating performance statistics and key metrics for each of the nine factors based on a Spearman rank information coefficient (IC) approach. For backtesting purposes, the factor portfolios are each rebalanced monthly over a 30-year time horizon using a rolling-window procedure.

Yuen and Ruckey consider a variety of metrics to assess the results of the factor portfolio backtests. Yuen asks Ruckey what can be concluded from the data for three of the factor strategies in Exhibit 1:

Exhibit 1 Backtest Metrics for Factor Strategies

	Factor 1	Factor 2	Factor 3
Thirty-year average signal autocorrelation	90%	80%	85%
Spearman rank IC: Month 1 (IC_1)	5%	4%	3%
Spearman rank IC decay speed	Fast	Slow	Modest

Yuen and Ruckey then run multifactor model backtests by combining the factor portfolios into two factor-weighted multifactor portfolios: an equally weighted benchmark portfolio (Portfolio A) and a risk parity portfolio (Portfolio B). Ruckey tells Yuen the following:

Statement 1 A risk parity multifactor model is constructed by equally weighting the risk contribution of each factor.

Statement 2 The process of creating Portfolios A and B requires a second rolling-window procedure in order to avoid model selection bias.

To gain a more complete picture of the investment strategy performance, Yuen and Ruckey design and then run two simulation methods to generate investment performance data for the underlying factor portfolios, assuming 1,000 simulation trials for each approach:

Approach 1 Historical simulation

Approach 2 Monte Carlo simulation

Yuen and Ruckey discuss the differences between the two approaches and then design the simulations, making key decisions at various steps. During the process, Yuen expresses a number of concerns:

Concern 1: Returns from six of the nine factors are clearly correlated.

Concern 2: The distribution of Factor 1 returns exhibits excess kurtosis and negative skewness.

Concern 3: The number of simulations needed for Approach 1 is larger than the size of the historical dataset.

For each approach, Yuen and Ruckey run 1,000 trials to obtain 1,000 returns for Portfolios A and B. To help understand the effect of the skewness and excess kurtosis observed in the Factor 1 returns on the performance of Portfolios A and B, Ruckey suggests simulating an additional 1,000 factor returns using a multivariate skewed Student's *t*-distribution, then repeating the Approach 2 simulation.

1 Following Ruckey's adjustments to the initial vendor database, backtested returns will *most likely* be subject to:

A stale data.

B look-ahead bias.

C survivorship bias.

2 Based on Exhibit 1, Ruckey should conclude that:

A Factor Strategy 1 portfolios experience the highest turnover.

B Factor Strategy 2 provides the strongest predictive power in the long term.

C Factor Strategy 3 provides the strongest predictive power in the first month.

3 Which of Ruckey's statements about constructing multifactor portfolios is correct?

A Only Statement 1

B Only Statement 2

C Both Statement 1 and Statement 2

4 Approach 1 differs from Approach 2 in that:

A it is deterministic.

B a functional form of the statistical distribution for each decision variable needs to be specified.

C it assumes that sampling the returns from the actual data provides sufficient guidance about future asset returns.

5 To address Concern 1 when designing Approach 2, Yuen should:

A model each factor or asset on a standalone basis.

B calculate the 15 covariance matrix elements needed to calibrate the model.

C simulate future factor returns using a joint cumulative probability distribution function.

6 Based on Concern 2, the Factor 1 strategy is *most likely* to:

A be favored by risk-averse investors.

B generate surprises in the form of negative returns.

C have return data that line up tightly around a trend line.

7 To address Concern 3 when designing Approach 1, Yuen should:

A bootstrap additional returns using a walk-forward framework.

B randomly sample from the historical returns with replacement.

C choose the multivariate normal distribution as the initial functional form.

8 The process Ruckey suggests to better understand how the performance of Portfolios A and B using Approach 2 is affected by the distribution of Factor 1 returns is *best* described as:

A data snooping.

B sensitivity analysis.

C inverse transformation.

The following information relates to Questions 9–16

Kata Rom is an equity analyst working for Gimingham Wealth Partners (GWP), a large investment advisory company. Rom meets with Goran Galic, a Canadian private wealth client, to explain investment strategies used by GWP to generate portfolio alpha for its clients.

Rom describes how GWP creates relevant investment strategies and then explains GWP's backtesting process. Rom notes the following:

Statement 1 Using historical data, backtesting approximates a real-life investment process to illustrate the risk–return tradeoff of a particular proposed investment strategy.

Statement 2 Backtesting is used almost exclusively by quantitative investment managers and rarely by fundamental investment managers, who are more concerned with information such as forward estimates of company earnings, macroeconomic factors, and intrinsic values.

Rom states that GWP is recognized in the Canadian investment industry as a value manager and that it uses traditional value parameters to build and backtest portfolios designed to outperform benchmarks stipulated in each client's investment policy statement. Galic, who is 62 years old, decides to allocate C$2 million (representing 10% of his net worth) to an account with GWP and stipulates that portfolio assets be restricted exclusively to domestic securities. Rom creates Value Portfolio I for Galic based on value factors analyzed in a series of backtests.

At a subsequent meeting with Galic, Rom explains the long–short hedged portfolio approach for implementing factor-based portfolios that GWP used to create Value Portfolio I and the steps involved in the backtesting procedure. One specific step in the process concerns Galic, who states the following:

Statement 3 I have never sold a stock that I did not own, and I really do not like the notion of giving the banks almost all of the income earned on the cash proceeds from the stock dispositions. On top of the forgone interest income, I think it could be really difficult to avoid high turnover and transaction costs, which would also negatively affect my risk-adjusted performance.

In an effort to relieve the concern raised by Galic, Rom suggests using an alternative backtesting approach to evaluate Value Portfolio I. This method uses the correlation between the prior-period ranked factor scores and the ranked current-period returns to evaluate the model's effectiveness. The approach generates a measure of the predictive power of a given factor relative to future stock returns.

Rom explains that the two backtesting approaches discussed so far have a weakness embedded in them. The approaches generally do not capture the dynamic nature of financial markets and in particular may not capture extreme downside risk. In an attempt to remedy this issue, Rom suggests considering different methods of modeling randomness. Rom states that GWP recently performed a statistical study of value and momentum factors, which found that both distributions were negatively skewed with fat tails. Additionally, the joint distribution of the returns for the factors had two peaks in the tails, and the peaks were higher than that from a normal distribution.

The study also investigated the return distributions of a number of individual value and momentum factors and found that they were non-normal based on their negative skewness, excess kurtosis, and tail dependence. Rom indicated that investment strategies based on this type of data are prone to significantly higher downside risk. Exhibit 1 compares downside risk measures for three model factors.

Exhibit 1 Downside Risk Measures for Model Factors

Risk Measure	Factor 1	Factor 2	Factor 3
Value at risk (VaR) (95%)	(6.49%)	(0.77%)	(2.40%)
Conditional VaR (CVaR) (95%)	(15.73%)	(4.21%)	(3.24%)
Maximum drawdown	35.10%	38.83%	45.98%

Rom explains that many of the examples used so far have incorporated the rolling-window backtesting approach. When comparing GWP's studies with those performed by Fastlane Wealth Managers for the same data and factors, Rom finds that the results are quite different. Rom discovers that Fastlane uses various modeling techniques, backtests each of them, and then picks the best-performing models. This discovery leads Rom to believe that Fastlane's modeling approach may exhibit selection bias.

Finally, after evaluating financial data that has periods of structural breaks, Rom informs Galic that GWP uses a technique commonly referred to as scenario analysis. This technique helps investment managers understand the performance of an investment strategy in different structural regimes. Exhibit 2 compares the performance of two factor allocation strategies under varying macroeconomic conditions.

Exhibit 2 Scenario Analysis Using the Sharpe Ratio

Strategy/Regime	High Volatility	Low Volatility	Recession	Non-recession
Strategy I	0.88	0.64	0.20	1.00
Strategy II	1.56	1.60	1.76	1.52

9 Which of Rom's statements concerning backtesting is correct?

A Only Statement 1

B Only Statement 2

C Both Statement 1 and Statement 2

10 The key parameter *most likely* to be incorporated in the analysis of Value Portfolio I is:

A monthly rebalancing.

B the MSCI World equity index.

C hedged returns into domestic currency.

11 In Statement 3, Galic expresses the *most* concern about the backtesting step that involves:

A strategy design.

B analysis of backtesting output.

C historical investment simulation.

12 The alternative approach to evaluate the backtesting of Value Portfolio I suggested by Rom is *most likely*:

A the Pearson information coefficient.

B the Spearman rank information coefficient.

C a cross-sectional regression.

13 Based on the statistical study performed by GWP, the tail dependence coefficient is *most likely*:

A low and negative.

B high and negative.

C high and positive.

14 Based on Exhibit 1, the factor with the smallest downside risk as measured by the weighted average of all losses that exceed a threshold is:

A Factor 1.

B Factor 2.

C Factor 3

15 The approach used by Fastlane Wealth Managers *most likely* incorporates:

A risk parity.

B data snooping.

C cross-validation.

16 Comparing the two strategies in Exhibit 2, the *best* risk-adjusted performance is demonstrated by:

A Strategy II in periods of low volatility and recession.

B Strategy I in periods of high volatility and non-recession.

C Strategy II in periods of high volatility and non-recession.

SOLUTIONS

1 A is correct. The analyst assumes a reporting lag of four months, which can introduce stale information even though it can significantly reduce look-ahead bias. C is incorrect because Ruckey has accounted for survivorship bias in back-testing by using point-in-time index constituent stocks, and not just the current survivors.

2 B is correct. Factor Strategy 2's signal provides the strongest predictive power in the long term because the Spearman rank IC for the first month is positive and the decay speed is slow.

A is incorrect because Factor Strategy 1 has the highest 30-year average signal autocorrelation of the three strategies, which indicates that portfolios formed using this factor experience the lowest turnover. All else being equal, factors with low turnover, indicated by a high autocorrelation, are preferred because such factors lead to lower portfolio turnover, lower transaction costs, and therefore higher after-cost cumulative performance.

C is incorrect because Factor Strategy 3 has the lowest Spearman rank IC for Month 1, indicating the weakest predictive power in the first month. The higher the average IC, the higher the factor's predictive power for subsequent returns.

3 A is correct. A risk parity multifactor model is constructed by equally weighting the risk contribution of each factor.

B is incorrect because Statement 2 is incorrect. The process for creating multifactor portfolios by equally weighting all factors and equal-weighting the risk contribution of all factors requires a second rolling-window procedure in order to avoid look-ahead bias, not model selection bias.

C is incorrect because Statement 1 is correct but Statement 2 is incorrect. The process for creating multifactor portfolios by equally weighting all factors and equal-weighting the risk contribution of all factors requires a second rolling-window procedure in order to avoid look-ahead bias, not model selection bias.

4 C is correct. Approach 1 is a historical simulation and assumes that past asset returns provide sufficient guidance about future asset returns.

A is incorrect because both approaches are non-deterministic and random in nature. Approach 1 is a historical simulation and Approach 2 is a Monte Carlo simulation.

B is incorrect because Approach 1 is a historical simulation and each random variable of interest (key driver and/or decision variable) is randomly drawn from historical data. A functional form of the statistical distribution of returns for each decision variable needs to be specified for a Monte Carlo simulation, which is Approach 2.

5 C is correct. Approach 2 is a Monte Carlo simulation. The returns of Portfolios A and B are driven by the returns of the nine underlying factor portfolios (based on nine common growth factors). In the case of asset or factor allocation strategies, the returns from six of the nine factors are clearly correlated, and therefore it is necessary to specify a multivariate distribution rather than modeling each factor or asset on a standalone basis. In the context of a multivariate normal distribution with nine random variables, nine randomly generated numbers from the uniform distribution are mapped onto a point on the joint cumulative probability distribution function.

A is incorrect because Approach 2 is a Monte Carlo simulation to generate investment performance data for the nine underlying factor portfolios. The returns of six of the nine factors are clearly correlated, which means specifying a multivariate distribution rather than modeling each factor or asset on a standalone basis.

B is incorrect because the analyst should calculate the elements of the covariance matrix for all factors, not only the correlated factors. Doing so entails calculating 36, not 15, elements of the covariance matrix. Approach 2 is a Monte Carlo simulation using the factor allocation strategies for Portfolios A and B for the nine factor portfolios, the returns of which are clearly correlated, which means specifying a multivariate distribution. To calibrate the model, a few key parameters need to be calculated: the mean, the standard deviation, and the covariance matrix. For 9 assets, we need to estimate 9 mean returns, 9 standard deviations, and $\frac{9 \times (9-1)}{2} = 36$ elements of the covariance matrix.

Assuming just the 6 correlated assets, the calculation is: $\frac{6 \times (6-1)}{2} = 15$.

6 B is correct. The distribution of Factor 1 returns exhibits excess kurtosis and negative skewness. The excess kurtosis implies that these strategies are more likely to generate surprises, meaning extreme returns, whereas the negative skewness suggests those surprises are more likely to be negative (than positive).

A is incorrect because risk-averse investors are more likely to prefer distribution properties such as positive skew (higher probability of positive returns) and lower to moderate kurtosis (lower probability of extreme negative surprises). The distribution of Factor 1 returns exhibits excess kurtosis and negative skewness.

C is incorrect because the distribution of Factor 1 returns exhibits excess kurtosis and negative skewness. The joint distribution of such returns is rarely multivariate normal—so, typically the means and variances of these returns and the correlations between them are insufficient to describe the joint return distribution. In other words, the return data do not line up tightly around a trend line because of fat tails and outliers.

7 B is correct. Random sampling with replacement, also known as bootstrapping, is often used in historical simulations because the number of simulations needed is often larger than the size of the historical dataset. Because Approach 1 is a historical simulation and Concern 3 notes that the number of simulations needed is larger than the size of the historical dataset, bootstrapping should be used.

A is incorrect because although bootstrapping (random sampling with replacement) would address the concern that the number of simulations needed is larger than the size of the historical dataset, a walk-forward framework is used for backtesting as a proxy for actual investing, not for bootstrapping. For historical simulation, researchers typically use a rolling window (also called walk-forward) framework to rebalance the portfolio periodically and then track the performance over time.

C is incorrect because choosing the multivariate normal distribution as the initial functional form is typically done in a Monte Carlo simulation (Approach 2), not in a historical simulation (Approach 1). Historical simulation randomly samples from the historical dataset by drawing a number from a uniform distribution so that there is equal probability of being selected. Choice of distribution would not address the concern about the size of the dataset.

8 B is correct. Sensitivity analysis can be implemented to help managers understand how the target variable (portfolio returns) and risk profiles are affected by changes in input variables. Approach 2 is a Monte Carlo simulation, and the results depend on whether the multivariate normal distribution is the correct functional form or a reasonable proxy for the true distribution. Because this information is almost never known, sensitivity analysis using a multivariate skewed Student's *t*-distribution helps to account for empirical properties such as the skewness and the excess kurtosis observed in the underlying factor return data.

A is incorrect because Ruckey is describing sensitivity analysis, not data snooping. Data snooping is the subconscious or conscious manipulation of data in a way that produces a statistically significant result (i.e., a *p*-value that is sufficiently small or a *t*-statistic that is sufficiently large to indicate statistically significance).

C is incorrect because Ruckey is describing sensitivity analysis, not inverse transformation. The inverse transformation method is the process of converting a randomly generated number into a simulated value of a random variable.

9 A is correct. Statement 1 is correct because the main objective of backtesting is to understand the risk–return tradeoff of an investment strategy by approximating the real-life investment process.

B is incorrect because Statement 2 is not accurate. Although backtesting fits quantitative and systematic investment styles more naturally, it has also been heavily used by fundamental managers.

C is incorrect because Statement 2 is not accurate. Backtesting is used in quantitative and systematic investment styles and is also heavily used by fundamental managers.

10 A is correct. A number of key parameters need to be specified when backtesting an investment strategy, including the investment universe, stock returns, frequency of rebalancing, and start and end dates. Because Galic has specified his desire to use only domestic (Canadian) equities, the MSCI World equity index and currency hedging of equity returns should not be used. Therefore, a key parameter to be incorporated into the analysis of Value Portfolio I is the frequency of rebalancing. Practitioners typically use monthly returns and monthly rebalancing.

11 B is correct. Backtesting typically follows the three steps of (1) strategy design, (2) historical investment simulation, and (3) analysis of backtesting output. Analysis of backtesting output encompasses the calculation of portfolio performance statistics and other key metrics such as turnover, the Sharpe ratio, the information ratio, and the Sortino ratio. Galic's apprehension about the long–short hedged portfolio approach relate primarily to issues (e.g., loss of interest income, transaction costs, turnover) that could negatively affect his risk-adjusted performance. Therefore, his statement expresses the most concern with matters measured in Step 3: analysis of backtesting output.

12 B is correct. The alternative approach to the long–short hedged portfolio method is to use an information coefficient (IC). The Spearman rank IC is the correlation between the prior-period ranked factor scores and the ranked current-period returns.

A is incorrect because the Pearson IC is the simple correlation coefficient between the factor scores for the prior period for all stocks in the investment universe under consideration and the current period's stock returns.

C is incorrect because, although a cross-sectional (univariate) regression approach may use returns at time t and factor scores at time $t - 1$, it does not use the correlation between the prior-period ranked factor scores and the ranked current-period returns, whereas the Spearman Rank IC does. A univariate regression's inference centers on whether or not the fitted factor return is statistically significant. Because the regression coefficient and the Pearson IC [i.e., Corr(r_t, f_{t-1})] always have the same sign, they typically produce similar results.

13 C is correct. The statistical study performed by GWP on the relationship between value and momentum factors found that the joint distribution between the returns for the factors had two peaks in the tails and that the peaks were higher than that from a normal distribution. This result implies a higher probability of positive co-movements in the tails of the two variables, relative to that from a normal distribution. Because the tail dependence coefficient focuses on the correlation in the tails of two random variables, the tail dependence coefficient in the study is most likely high and positive.

14 C is correct. Exhibit 1 presents three downside risk measures: VaR, CVaR, and maximum drawdown. Conditional VaR is defined as the weighted average of all loss outcomes in the return distribution that exceed the VaR loss. Thus, CVaR is a more comprehensive measure of tail loss than VaR. Based on Exhibit 1, the factor with the smallest downside risk based on CVaR is Factor 3.

15 B is correct. The fact that the two firms' investment performance results differ over similar time horizons using the same data and factors may be the result of selection bias. Data snooping is a type of selection bias. Fastlane Wealth Managers is most likely selecting the best-performing modeling approach and publishing its results.

A is incorrect because risk parity is a portfolio construction technique that accounts for the volatility of each factor and the correlations of returns among all factors to be combined in the portfolio. It is not regarded as selection bias.

C is incorrect because cross-validation is a technique used in the machine learning field to partition data for model training and testing. It is not considered selection bias.

16 A is correct. Using the Sharpe ratio, the best risk-adjusted relative performance can be determined by comparing the sensitivity of the two strategies under differing macroeconomic regimes: recession versus non-recession and high volatility versus low volatility. The best risk-adjusted return will exhibit the highest Sharpe ratio. Strategy II demonstrates higher risk-adjusted returns compared with Strategy I under all four macroeconomic conditions, particularly in periods of low volatility, when the Sharpe ratio outperformance is 0.96, and recessions, when the Sharpe ratio outperformance is 1.56.

Scenario Analysis Using Sharpe Ratio

Strategy/Regime	High Volatility	Low Volatility	Recession	Non-recession
Strategy I	0.88	0.64	0.20	1.00
Strategy II	1.56	1.60	1.76	1.52
Difference (II – I)	0.68	0.96	1.56	0.52