An Experiment with Objective Metrics

Sharpe and Information Ratio May Not Be Best

Source Code: <https://github.com/Vincent-Chin/Risk_Perspective_Paper>

# Summary

In active portfolio management, Sharpe Ratio and Information Ratio are commonly used as a criterion to infer the future quality of an investment strategy. This paper conducts experiments using a hypothetical active selection strategy to evaluate these ratios as predictors for out-of-sample performance, along with the Sortino and MAR Ratios.

Through these experiments, I observe that while both Sharpe and Information Ratio are interesting backward-looking metrics, both chronically underperform when used as portfolio predictors on a walk-forward basis. I find similar results in both traditional walk-forward testing and quasi-Monte Carlo walk-forward testing.

# Methodology

## Experiment Setup

This experiment takes place over all the US trading dates between 1/1/2014 and 12/31/2018. There are a total of 1304 trading dates in that timespan.

For each trading date, I re-evaluate each permutation (one permutation being a ‘potential’ strategy) of the search universe using a 100-day lookback. I adopt the best-performing permutation for each date. The trades belonging to that exact date I designate as our ‘in-sample’ trades, and the trades for that permutation that would occur on the following date I designate as our ‘out-of-sample’ trades. This corresponds to a fairly standard walk-forward optimization[[1]](#footnote-1).

I then analyze the aggregated performance of the in-sample trades and compare it against the aggregated performance against the out-of-sample trades. The behavior of interest is the contrast between the in-sample and out-of-sample performance when holding the objective metric constant.

I conduct this experiment twice: first, with a traditional linear evaluation – where I look at the last 100 trading days and assess their performance to choose a strategy permutation. And then second, with a quasi-Monte Carlo[[2]](#footnote-2) generated evaluation – where I assemble 50 random samplings of the last 100 trading days and average the performances of each sample, and then used those averaged performances to select a strategy permutation.

## Data Environment

To conduct this experiment, I selected the stocks making up the SP100 index as of 4/27/2019, except for: BRK.B, DOW, FOX, FOXA, due to difficulties obtaining data for those tickers. I also selected the SPY ETF to act as a general performance benchmark. For each of the chosen stocks, I obtained the daily adjusted-close (inclusive of dividends and splits) from 1/1/2014 to 12/31/2018 and used those adjusted-closes to generate returns calculations.

I noted that using a 2019 snapshot does not reflect the adds and removes from the SP100 index over time, so there is a degree of tracking error when using the SPY as a benchmark. I considered it an acceptable compromise given that our primary objective is about comparing performance metrics rather than the benchmark itself.

## Search Universe, Strategy Dimensions, and Position Accounting

For this experiment, I devised a simple long-only momentum strategy that uses two moving averages: the 20-day and the 200-day. For any given trading day, this strategy calculates the following value for both averages:

*The closing price’s distance from the n-day mean, normalized in terms of the n-day standard deviation.*

These values are generally bounded at -2.5 to +2.5 standard deviations. From here, I build a search universe where every permutation in the search universe exercises one possible filter.

In the search universe, I test every increment of 0.5 standard deviations from -2.5 to 2.5 for both the 20-day and 200-day averages, and I also test both permutations of the relationships themselves (greater than as a signal, and less than as a signal.) This generates a search universe of 400 possible total permutations.

For example, one combination might be:

*Enter long if: (20-day distance > 2.0 20-day standard deviations)*

*and (200-day distance < -1.5 200-day standard deviations)*

For any given day, this filter generates a subset of stocks in the universe that I can take a long position in. I allocate to each stock equally such that the overall return is a simple average of the component stocks’ returns. (i.e. if there are five stocks remaining after filtering, then the portfolio will be exactly 20% exposed to each stock.)

Positions are rebalanced daily; returns are calculated as close-to-close and assumed to happen without commissions.

These conditions are somewhat unrealistic in reality. In practice, by the time a computer system will have calculated the new position structure, the market will have already closed, making it impossible to obtain the prior closing price to initiate the new positions. It is also highly unlikely that a strategist is able to trade for free or able to perfectly allocate evenly between stocks. However, since I am only attempting to measure relative performance between several objective metrics for the same strategy, I did not consider these to be critical issues.

## Objective Metrics

I test four different objective metrics in this experiment, Sharpe Ratio, Sortino Ratio, MAR (Drawdown) Ratio, and Information Ratio.

The ratios are defined as follows:

Sharpe Ratio: annualized arithmetic return

/ annualized standard deviation of return

Sortino Ratio: annualized arithmetic return

/ annualized semi-standard deviation of return,

Where all returns > 0 are replaced by 0 when calculating the semi-standard deviation

MAR Ratio: annualized arithmetic return

/ greatest observed peak-to-trough drawdown

Info Ratio: annualized daily performance differences vs SPY benchmark

/ standard deviation of daily performance differences vs SPY benchmark

I assume a risk-free rate of 0 for all ratios, as the performance versus the benchmark itself is not central to the results of this experiment.

# Observations and Findings

The most obvious finding is that this moving average strategy is not a very good one. Under both experimental paradigms (traditional and Monte Carlo), the out-of-sample performance is mostly negative regardless of objective metric used. (See appendix 2 and 3.)

Of more interest, however, are Tables 1 and 2, which show us the performance differences between out-of-sample and in-sample in both absolute (table 1) and relative (table 2) terms. Here, I see that as objective metrics, both Sharpe and Information Ratio suffer from the largest percentage drop-offs on all metrics except for Information Ratio, denoted in the column ‘Avg Loss (exInfo)’. This is uniform for both the traditional test and the Monte Carlo test.

The MAR Ratio, by contrast, seems to lose the least – suffering a ~15% smaller performance hit in the traditional method and ~53% smaller performance hit in the Monte Carlo method, as compared to the Information Ratio. The Sortino ratio has similar (though slightly worse) performance.

As an interesting aside, the absolute drop-offs for all objective metrics are much smaller when using the Monte Carlo method. This would imply that the Monte Carlo approach that I’ve employed is much more trustworthy than traditional walk-forward simulation, on a forward-looking basis.

Appendix 4 shows this visually. In traditional walk-forward, our in-sample performance massively outperforms the SPY benchmark (in gray), but of our corresponding out-of-sample performances (in pastels), only MAR manages to roughly even match the benchmark. The rest come in below.

Appendix 5 shows that by contrast, using the Monte Carlo method I have a much more similar performance between in-sample and out-of-sample methods, to the extent that it is unclear which are in-sample and which are out-of-sample without first looking at the key.

The exemplary in-sample performance numbers from the traditional walk-forward simulation are very misleading, considering how similar the out-of-sample performance numbers are for both methods.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 1: Performance Difference of Out-of-Sample vs. In-Sample, Absolute Differences** | | | | | | | |
| **Perf - Absolute** | **Walk Method** | **Sharpe** | **Sortino** | **MAR** | **InfoRatio** | **Avg Loss (exInfo)** | **Avg Loss** |
| Sharpe | Traditional | (2.42) | (4.31) | (2.69) | (2.36) | (3.14) | (2.94) |
| Sortino | Traditional | (2.08) | (4.60) | (3.49) | (1.84) | (3.39) | (3.00) |
| MAR | Traditional | (1.46) | (3.44) | (3.72) | (1.23) | (2.87) | (2.46) |
| InfoRatio | Traditional | (2.50) | (4.47) | (2.66) | (3.94) | (3.21) | (3.39) |
| Sharpe | Monte Carlo | (0.19) | (0.27) | (0.12) | (0.27) | (0.19) | (0.21) |
| Sortino | Monte Carlo | (0.45) | (0.67) | (0.21) | (0.39) | (0.44) | (0.43) |
| MAR | Monte Carlo | (0.17) | (0.25) | (0.08) | (0.14) | (0.16) | (0.16) |
| InfoRatio | Monte Carlo | (0.19) | (0.28) | (0.06) | 0.16 | (0.17) | (0.09) |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 2: Performance Difference of Out-of-Sample vs. In-Sample, Normalized to In-Sample Basis** | | | | | | | |
| **Perf - Relative** | **Walk Method** | **Sharpe** | **Sortino** | **MAR** | **InfoRatio** | **Avg Loss (exInfo)** | **Avg Loss** |
| Sharpe | Traditional | -105% | -104% | -103% | -136% | -104% | -112% |
| Sortino | Traditional | -95% | -97% | -98% | -116% | -97% | -101% |
| MAR | Traditional | -82% | -87% | -95% | -100% | -88% | -91% |
| InfoRatio | Traditional | -106% | -105% | -102% | -122% | -104% | -109% |
| Sharpe | Monte Carlo | -203% | -197% | -331% | -85% | -243% | -204% |
| Sortino | Monte Carlo | -177% | -174% | -182% | -426% | -178% | -240% |
| MAR | Monte Carlo | -179% | -175% | -183% | -75% | -179% | -153% |
| InfoRatio | Monte Carlo | -257% | -251% | -190% | 23% | -232% | -169% |

# Potential Explanations and Opportunities for Further Research

I take it for granted that on a backwards-looking perspective (that is, on things that have already happened), all four metrics that I’ve tested are useful. However, why does the performance gap exist for Sharpe and Information when trying to use them as forward-looking objective metrics?

Intuitively, I’d seek to blame Information Ratio’s performance gap on recency effect, or on the phenomenon of chasing winners. With Information Ratio, we are not gauging performance of the strategy against itself, but rather on its performance relative to an external benchmark. This creates contingencies due to the idiosyncratic behavior of the benchmark – you could’ve gone short while the benchmark happened to be in an extended down-turn, for example. Then, all permutations of “short” would rise to the top as strategy candidates, even though for subsequent periods the downturn may no longer be there.

The observation that Information Ratio suffers less percentage-wise in the Monte Carlo simulation than it does in the traditional simulation would lend some credence to this hypothesis. Since the Monte Carlo method randomly samples from the testing period, it minimizes any extended idiosyncratic trends of the benchmark that would distort performance in future periods. The traditional simulation method does not have this defense, and the out-of-sample Information Ratio decreases accordingly.

As for why Sharpe underperforms – that may be due to the nature of the metric itself. The stock market is widely thought to be asymmetrically slanted towards the upside – there is a slight tilt towards positive returns more than negative returns. It wouldn’t be unreasonable to conject that this corresponds to outliers as well.

The Sharpe Ratio, however, penalizes equally for both positive and negative outliers. Then, when used as an objective metric, it would naturally slant towards the most conservative strategies that avoid these outliers entirely.

Then, in out-of-sample trading, these “lucky” wins would have very little chance of happening – at least in comparison to the MAR and Sortino ratios, which do not penalize against upside outliers – which would lead to Sharpe’s underperformance accordingly.

Alternatively, this phenomenon for Sharpe may only be at play because of the nature of the strategy. In this experiment, I’ve used a momentum strategy; momentum strategies exhibit negative skew[[3]](#footnote-3) in that they have mostly small winners with a few large losers. MAR and Sortino ratios would naturally flourish for such distributions, because even-more-negatively-skewed distributions that have larger winners would not affect the denominators for those two measures, while they would impact the denominator for the Sharpe ratio, and thus be more likely to be eliminated from consideration.

By contrast, if I’d used a mean-reversion strategy with a positive skew (mostly small losers and a few large winners), something different might happen – even-more-positively-skewed distributions with larger losers would affect Sharpe, MAR, and Sortino with the same level of impact, potentially removing the performance difference. This would need to be proven with a different experiment.

# Conclusion

In this experiment, I used two methods to compare out-of-sample and in-sample performance for several popular ratios: Sharpe, Sortino, MAR, Information. My experiments show that for both methods, Sharpe and Information Ratio underperform as forward-looking criteria, at least in-comparison to Sortino and MAR.

I suggest that this is first because Information Ratio may suffer from recency effects or a “chasing winners” phenomenon, and second because Sortino and MAR do not penalize on upside volatility, while both Sharpe and Information Ratio do.

However, it is possible that the deficient performance observed in Sharpe and Information Ratio are a consequence of the distribution skew of the strategy itself – there isn’t enough information from the experiments to conclude this definitively.

The real answer may be that there is no one-size-fits-all metric that can compare between all strategies. Instead, it might be that for certain types of strategies we should use MAR and Sortino, and for other types of strategies we should use Sharpe and Information Ratio instead.

# Appendix 1: Stock Universe and Strategy Parameters

# https://en.wikipedia.org/wiki/S%26P\_100 - snapshot 4/27/2019

# excludes (data retrieval failure): 'BRK.B', 'DOW', 'FOX', 'FOXA',

symbols = ['AAPL', 'ABBV', 'ABT', 'ACN', 'AGN', 'AIG', 'ALL', 'AMGN', 'AMZN', 'AXP',

'BA', 'BAC', 'BIIB', 'BK', 'BKNG', 'BLK', 'BMY', 'C', 'CAT',

'CELG', 'CHTR', 'CL', 'CMCSA', 'COF', 'COP', 'COST', 'CSCO', 'CVX', 'DHR',

'DIS', 'DUK', 'DWDP', 'EMR', 'EXC', 'F', 'FB', 'FDX',

'GD', 'GE', 'GILD', 'GM', 'GOOG', 'GOOGL', 'GS', 'HD', 'HON',

'IBM', 'INTC', 'JNJ', 'JPM', 'KHC', 'KMI', 'KO', 'LLY', 'LMT', 'LOW',

'MA', 'MCD', 'MDLZ', 'MDT', 'MET', 'MMM', 'MO', 'MRK', 'MS', 'MSFT',

'NEE', 'NFLX', 'NKE', 'NVDA', 'ORCL', 'OXY', 'PEP', 'PFE', 'PG', 'PM',

'PYPL', 'QCOM', 'RTN', 'SBUX', 'SLB', 'SO', 'SPG', 'T', 'TGT', 'TXN',

'UNH', 'UNP', 'UPS', 'USB', 'UTX', 'V', 'VZ', 'WBA', 'WFC', 'WMT',

'XOM'

]

initial\_date = "2014-01-01"

final\_date = "2018-12-31"

benchmark = 'SPY'

search\_parameters = {

'MA20\_Std': np.arange(-2.5, 2.5, 0.5),

'MA200\_Std': np.arange(-2.5, 2.5, 0.5),

'MA20\_Mode': ('>', '<'),

'MA200\_Mode': ('>', '<'),

}

objective\_metrics = ['Sharpe', 'Sortino', 'MAR', 'InfoRatio']

wf\_lookback = 100

mc\_samples = 50

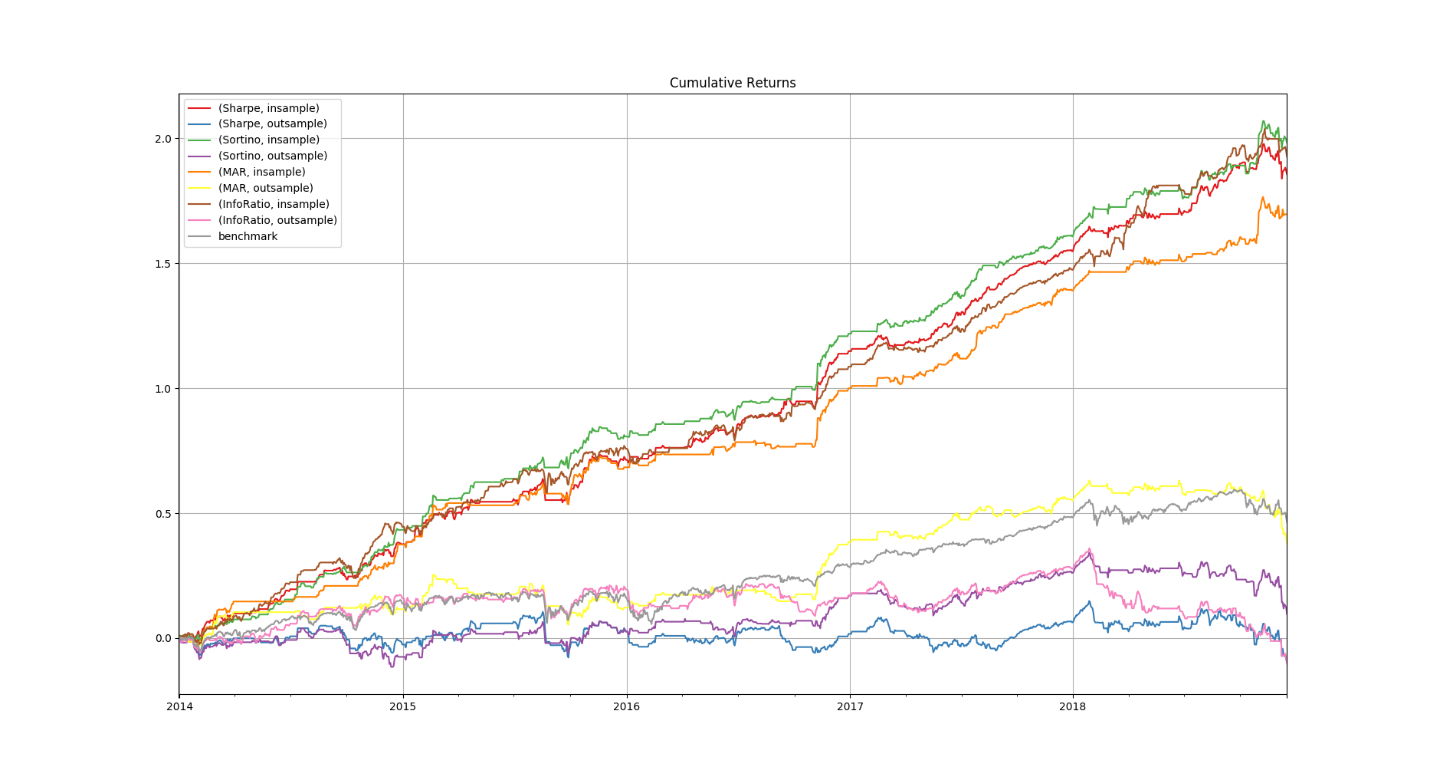
# Appendix 2: Traditional Walk-Forward Performance



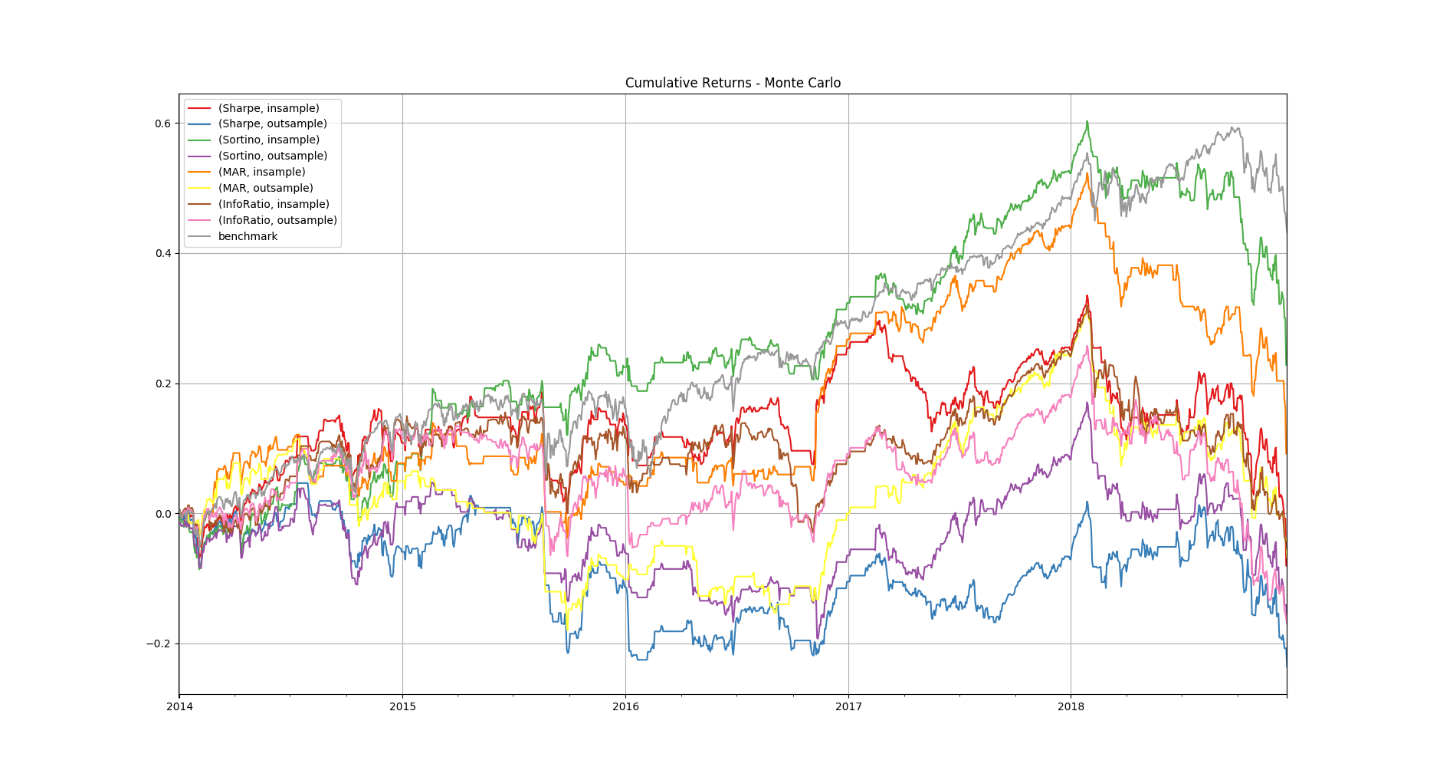
# Appendix 3: Monte Carlo Walk-Forward Performance



# Appendix 4: Cumulative Profit-Loss, Traditional Walk-Forward

****

# Appendix 5: Cumulative Profit-Loss, Monte Carlo Walk-Forward

****

1. https://en.wikipedia.org/wiki/Walk\_forward\_optimization [↑](#footnote-ref-1)
2. https://en.wikipedia.org/wiki/Quasi-Monte\_Carlo\_method [↑](#footnote-ref-2)
3. “Skewness Preference and Price Momentum.” http://myacme.org/IJMTP/issue4\_2003/2003\_paper6.pdf [↑](#footnote-ref-3)