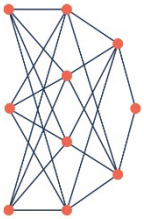


BACKTESTING IS ABOUT THE PAST

June 2022

I see you looking for alphas, man, like there's a bunch.

You're maybe trying to make money, ok, I really don't judge.

But if markets are efficient, trust me, you can't do much.

'Cause that's something we all know, there's no free lunch.

You can use your experience to follow the signs.

History does not repeat itself, but it often rhymes.

Some managers may tell you lots of pretty lies.

About how their portfolios once were butterflies.

Look, I'm not saying past performance is useless.

I'm sure we can use math to prove its worthiness.

But be careful with people trying to sell you just success.

Backtests are not about the future, yo, they're all about the past.

- Hully Rolemberg

Hully Rolemberg
Partner

Gabriel Koike
Partner

We also thank Erik Buischi and Paulo Hermanny for valuable comments and suggestions. Daemon Investments may or may not apply similar investment techniques or methods of analysis as described herein. The views expressed here are those of the authors and not necessarily those of Daemon Investments.

Contents

Introduction

We hope you have enjoyed this rapped introduction, backtesting is so important in our quant life that it deserves such an elegant approach. In fact, "backtest" may be one of the most used words in the quantitative industry and it simply means "testing some investment strategy using historical data". In other words, it simulates how a strategy would have performed ex-post, that is, had it been deployed in the past and had the past unfolded exactly as it did. Among other results, the backtest computes daily profit and loss that such strategy would have generated, which can be summarized in popular performance metrics like Sharpe ratio and maximum drawdown, for example. It sounds exceedingly comforting - but as with most things, the devil is in the details.

backtested performance should be considered a necessary but not sufficient condition. In this article, we will highlight 5 important aspects of backtesting that sometimes are taken for granted, but first let us see how a backtest works.

The name speaks for itself: "back" because it uses past data, and "test" because it is used to test a trading system. There are different ways to build a backtest depending on the kind of strategy we are trying to evaluate, but the general structure is as described in Figure 1.

First, we select the data. Backtests usually use public market data such as prices, volumes, and multiples. Data is processed (e.g., returns calculation), filtered (e.g., only assets from a market index are selected), and then inputted

Figure 1

Backtesting can be used to assess the viability of investment strategies, to guide investment decisions, and to decide optimal parameters combinations (e.g., rebalancing rules, stop losses, risk limits, etc.). However, it must be interpreted very carefully since it relies on non-trivial assumptions and violations of these conditions can generate misleading results that completely invalidate the analysis. When assessing investment strategies, good

in the trading system algorithm. The trading system defines a sequence of rules that represent the strategy we want to backtest subject to a set of assumptions (remember that models are simplified versions of reality that on rely on specific assumptions).

Suppose we run a vol-scaled long-only strategy following the simplified backtesting workflow detailed above. Signal calculation is simple: buy

all assets. Then, the system calculates assets' weights on the portfolio depending on its volatility so that the higher the volatility the smaller the position. The final step is to establish the position sizing by leveraging all the positions in accordance to some rule of capital distribution - this rule could be to invest all assets, in a long only portfolio, or to meet some volatility based criterion.

The outcome of this backtesting procedure is a historic series representing the performance of our trading system over time. We can use this result to calculate performance metrics, compare it to some benchmark, or even other strategies, and decide whether it is a good trade or not.

#1 History is just one realization of a stochastic process

From a statistical perspective, everything we observe in the world can be seen as just one realization of some unknown stochastic process. In other words, history could have been completely different in an unexpected way simply because it is just one realization of the generating process. However, even though we cannot know every possible outcome (since we do not know the true probability distribution), we can stress our backtests to some extent.

Unlike other types of data, when we use financial returns we have to be careful with any kind of perturbation we introduce in the series because financial returns have their own set of stylized facts that must be respected. For example, most assets exhibit returns that are weakly stationary (constant mean and constant autocovariance throughout time, finite second moment) by definition so we cannot introduce non-stationary perturbations (e.g., a deterministic trend) and apply the same estimation framework used to evaluate financial returns. The same is true for bootstrapping and Monte Carlo simulation approaches, for example.

Bootstrapped Returns

Bootstrapping is just a fancy word for *resampling,* and it is especially useful when the sample size is small. We use bootstrap techniques to generate new data from our original observations, that is, instead of having just one sample of realized returns we can have thousands of them using the original observations sampled in different ways. However, most of bootstrap approaches assume nothing about the distribution of data, what can generate new samples that do not necessarily relate to the original one.

A possible solution for this problem is the Stationary Bootstrap, introduced by Politis and Romano (1994), which is a procedure based on resampling “blocks of blocks” of consecutive observations (with random length) to form a pseudo-time series such that the original statistical properties (e.g., stationarity) are fully preserved.

Monte Carlo Simulation

While *bootstrapping* is a very interesting test for statistical robustness, it is restricted to resampling from one single realization of the process. Monte Carlo simulations are a class of simulations that intends to generate alternative realizations of a random process and evaluate its effects. This can be achieved by generating several alternative “histories” of the process, such as simulating random returns, or by

corrupting existing returns with a random process. Monte Carlo simulations are extensively used on option pricing models and on strategies that rely on said models.

One of the greatest challenges of Monte Carlo simulations is the set of assumptions embedded on the simulation. Since a Monte Carlo simulation relies on random processes as simulated market data as inputs to the model, it is highly subject to the "Garbage In Garbage Out" rule (described in the next section). Random process generation relies on probability distributions, assumptions about autocorrelation and correlation between different assets, for example. Even disturbing the market data with a non-correlated random process could change the structure on which a strategy relies on.

#2 Garbage in, garbage out.

Survivorship Bias

Data are the sole input of any backtest analysis, and it must be the best representation of a particular historic period of time, that is, it must represent the conditions of that time with high fidelity. A very common backtesting pitfall is to ignore assets that no longer exist, we call it *survivorship bias.* The straightforward example is a strategy that replicates an index but uses its latest composition and/or does not account for dropped assets.

There is a huge list of assets that are not traded in the present, but were traded someday in the past, just think of a stock that has been incorporated by another or even some discontinued ticker that went bankrupt. If historic data does not include these assets, then we will be introducing survivorship bias and our backtest results will not represent how that strategy would really have performed in the past.

For example, we all know that the S&P 500 is dynamic. The requirements for a company to be included in the index are (1) market cap larger than US$13.1 billion, (2) positive earnings, (3) the majority of its shares must be in the hands of public, and (4) it must be approved by an S&P Down Jones Indices Committee. That said, the composition S&P 500 might change in a non-trivial fashion over time. For example, the current index is heavily tilted towards tech industry, but it certainly was not the case 20 years ago.

Suppose you are running a backtest using data from the past 20 years and use the current composition of S&P 500 to select the assets. Some of these tech companies did not even exist back in 2000s: Facebook (now "Meta Plataforms", META) is currently the 12th largest position in the index, but the company was created in 2004, went public in 2012 and joined the index just in 2013. So, if you ran a backtest including META, (1) it would only have 10 years of historical data on this asset and (2) between 2012 and 2013 it would not represent S&P 500 assets. Besides that, current S&P companies can be a very misleading list of relevant companies 20 years ago, so we would be ignoring important companies from sectors such as industrials and utilities that used to be relatively more important in the past than they are now.

Revised Data

Similarly to the survivorship bias, another common backtesting error is using revised data instead of point-in-time data. Financial

information can be revised multiple of times once it is released: financial statements can have their previously released values adjusted during the next quarterly release (the industry calls this a *restatement*). Restatements are quite common in the financial industry but using restated data in a backtest is essentially using data gathered form the future – an immediate issue since the future is unknown. Still, that is an error that can easily slip by.

Restatements are necessary when it is determined that a previous statement contained a "material" inaccuracy, which can be resulting from fraud or simply accounting mistakes. Suppose a company restates its financial statements for 2016 and 2017 after auditors discover accounting errors for income taxes and that the understated income tax expense boosted net profits by US$400 million in 2016. If you were running a strategy based on the value factor at that time, you would have increased your exposure to this company, even though these earnings were completely artificial.

This episode did happen in 2019 and the company was Molson Coors Brewing Co. On the day after restatement, the price of shares in Molson Coors fell 9.4 percent, as the finding did not inspire much confidence in company's accounting practices. In this case, a backtest that does not use point-in-time data would produce unreliable results.

Corporate Events and Adjustments

Stocks (and stock derivatives) are subject to corporate events, such as dividends, splits, mergers, spin-offs and others. These need to be carefully accounted for, since they may create discontinuities in price series. When creating a series of returns, for example, these discontinuities must be compensated, otherwise the price data will appear to create unusually high (or low) "returns" arising from these corporate events.

And while for return series some of these adjustments are done and may create adjusted price series, for other purposes the actual point-in-time data may be necessary (such as when calculating multiples such as price-to-earnings, for example) - that is, not only these events need to be accounted for, but also each event needs to be correctly accounted for according to the intended use of the series.

Other types of adjustments can be done, for example, for series of future contracts. For example, suppose a strategy only trades on the first active contract of a future, such as in a currency hedge. Each different contract term has different prices, due to the futures term structure that incorporates carry-rolldown and other price expectation differences. For some applications, these outsized price discontinuities may pollute the signals we are trying to assess and as such we may wish to consider a "continuous" series for prices (i.e. adjusted for all the carry/roll effects). Then adjustments in prices are necessary to compensate to the differences in different contracts due to the futures term structure. These and other adjustments may also be necessary when creating return series for trading specific futures series.

#3 Overestimated Performance

Optimistic Transaction Costs

Backtests can be easily fine-tuned to generate beautiful results. In fact, one of the most

common rookie mistakes is to underestimate transaction costs, ultimately leading to an overestimated portfolio performance. There are three main types of transaction costs that we must acknowledge in our backtesting process: fees, slippage, and liquidity costs.

Obviously, trading is not costless, and this cost can actually be incremental: the more you trade, the more you pay. Overall, algo-traders (just like regular traders) depend on a brokerage intermediary to access exchanges, so these traders pay brokerage fees and commissions per trade to run their quantitative strategies. Besides that, government taxes can also be charged per trade, adding up to these incremental costs and eroding returns. Luckily, this kind of cost is usually fixed according to each asset/exchange/broker and well known ex-ante.

A not so obvious cost is the one that comes from the latency in trading, that is, the fact that you might not be able to trade at the price your backtest originally set, this difference in prices is called slippage. Observe that assets with higher volatility are more likely to show a larger slippage, but it also depends on the strategy that is being traded: high frequency, low frequency, trend following, etc.

Finally, there is the cost of liquidity, in other words, the impact of those trades on market's price. Liquidity costs depend on the asset liquidity versus the order size. Large orders on illiquid assets can move the market as they require a large part of current supply. Besides that, illiquid assets have larger bid-ask spreads, which is also a source of transaction costs since the algo-trader may not trade at any price close to the one defined ex-ante.

Multiple Testing

Ignoring biases introduced by multiple testing is perhaps one of the most dangerous mistakes one can make when evaluating backtested results – at least from a purely statistical standpoint. Suppose you developed a preliminary algorithmic trading system, and you want to backtest it to see how it would have performed in the past before moving on with your implementation. You run it for the first time and observe the model could be specified a little bit differently, so you change it and run the backtest again. The performance is better this time, but there is still room for improvement - you could use a different set of assets, for example. You keep doing this until the results seem appropriate (aka "the performance is good enough"). We call it multiple testing.

It is true that there is inevitable data mining by both the researcher and by other researchers in the past. So even if you run a single backtest with no iterations, you still have to account for biases emerged from multiple testing, since you may have been inspired by previous research that included multiple testing. A common solution for the multiple testing issue on backtests' performance is to discount the Sharpe ratio from a backtested portfolio. A rule of thumb is to discount the Sharpe ratio by 50% (or 25%), but we can use a statistical framework that systematically corrects Sharpe ratios by transforming it into a t-ratio.

Suppose that t-ratio is 3.0. While a t-ratio of 3.0 is highly significant in a single test, it might not be if we take multiple tests into account. We proceed to calculate a p-value that appropriately reflects multiple testing. We need to make an assumption on the number of

previous tests (trials). Suppose the adjusted p-value is 0.05. We then calculate adjusted t-ratio which, in this case, is 2.0. With this new t-ratio, we compute the "correct Sharpe ratio".

There are three main methodologies to adjust the p-value: Bonferroni, Holm, and BHY.

1. Bonferroni's method adjusts each p-value equally, it inflates the original p-value by the number of tests *M*.
2. Holm's method relies on the sequence of p-values (starting from the smallest) and corrects them similarly to Bonferroni's method. Both procedures are designed to eliminate all false discoveries no matter how many tests for a given significance level.
3. Benjamini, Hochberg and Yekutieli (BHY)'s procedure defines the adjusted p-values sequentially starting from the largest and defining the adjusted p-value sequence through pairwise comparisons.

The three procedures provide adjusted p-values that control for data mining, then we transform the corresponding t-ratios into Sharpe ratios.

#4 Forward-looking is dangerous

Theoretical vs Real Information Set

Ensuring the correct execution dynamics becomes critical for a backtest. Imagine you develop a trading system that uses closing prices to calculate positions. You must be aware that you only know closing prices when markets actually close. This means that even though in your backtest you might have closing prices from day *t* on the same day *t,* this is not true when you are trading live, so you must adjust your backtests' information set to be powered by the correct execution dynamics, otherwise you will be forward-looking data.

How dangerous really is forward-looking information to the reliability of backtests? Imagine that you use intraday prices to estimate volatility and you use this estimate to scale your portfolio's daily positions. Assume that the trading system sends orders to adjust your portfolio's positions according to the backtesting environment each day in the morning. In this scenario, you always forecast the upcoming volatility perfectly and your portfolio is always optimally scaled (given an intraday volatility estimator).

The same goes for establishing the desired positions. If the strategy decides assets to go long or short based on returns or ratios (such as value related ratios) that are available at the end of the day, the positions can only be traded in the following day. This may seem obvious, but the implementation can be tricky, because while most market data is end-of-day, some calculations are performed throughout the trading day (such as positions or trading P&L).

Another source of look-ahead bias is embedded in the construction of the trading strategy itself. A trading strategy is composed of selected market data, some trading logic - which can be induced from observation of data, and a systematic implementation of the logic - which eventually must contain quantitative parameters. All of these are subject to look-ahead bias. We described look-ahead biases in the context of data selection in Section #2. Suppose you are backtesting a stock strategy for the last 15 years and you select all companies that have been continuously traded within those years. This is problematic because it incorporates information of all periods into

the same period under evaluation, and includes survivorship bias, for example. Similarly, both logic and parameters need to be established with data available at the time, not using the whole information set. Some of this is discussed in the following section.

#5 Overfitting is lethal

Given any financial series, it is not difficult to overfit an investment performance, that is, to develop a trading model that targets particular observations (rather than a general structure) and performs pretty well in-sample (IS). What researchers and practitioners usually do is to split the sample into two subsets: the training set and the testing set. The training set is used to train the model and develop the trading strategy, whereas the testing set is used to run the trading model and evaluate its performance out-of-sample (OOS). If the performances both in-sample and out-of-sample are similar, then there is evidence to reject the hypothesis that the backtest is overfit. This is a simple and effective procedure, but it can easily fail.

If there is seasonality in data (which is not rare in financial series) then it is quite easy to know how variables performed over the testing set and this information ends up being used in modeling the strategy, consciously or not. Besides that, if the sample is small, the training set will be too short to fit, and the testing set too short to conclude anything with a certain level of confidence. Finally, depending on the time period selected to train versus to test the model, we may be led to opposite conclusions.

Observe that, if the testing set is taken from the end of the time series, we are not considering enough of the most recent observations (which are often the most representative ones) in the fitting sample. On the other hand, if the test set is taken from the beginning of the sample, the out-of-sample evaluation will be done on the least representative portion of data. Besides that, sample selection is something quite easy to manipulate in the direction that favors strategies' performance the most.

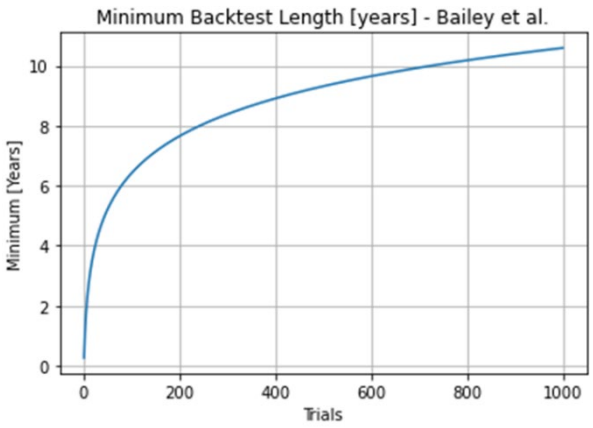
Selecting both IS and OOS samples is a problem that we must address very carefully because, as we said, it is pretty easy to overfit a backtest. Bailey and López de Prado (2012) show that The Minimum Backtest Length (MinBTL, in years) needed to avoid selecting a strategy with an IS annualized Sharpe Ration of $E[max_N]$, among *N* ibid-normally-distributed strategies with an expected OOS SR of zero is given by Equation 1, where $\boldsymbol{\gamma}$ is the Euler-Mascheroni constant, Z is the CDF of the Standard Normal, and *e* is Euler's number. The minimum backtest length (in years) as a function of the number of trials is represented in the Figure 2.

$$MinBTL \approx \left(\frac{(1-\gamma)Z^{-1}\left[1-\frac{1}{N}\right] + \gamma Z^{-1}\left[1-\frac{1}{N}e^{-i}\right]}{E[max_N]} \right)^2 < \frac{2\ln(N)}{E[max_N]^2}$$

Equation 1

If only 5 years of data are available, no more than 45 independent model configurations should be tried, and for this number of trials, the expected maximum Sharpe ratio in sample

is 1, whereas the expected Sharpe ratio out-of-sample is 0 (see Bailey et al 2014 and Bailey et al 2015). That said, if we do not know the number



of trials (or at least an estimative of it) used to obtain the selected backtest configuration, it is impossible to assess the risk of overfitting and these backtested results cannot be a reliable source of the trading system's performance.

Ok, but why is overfitting lethal?

Remember that we may choose a particular parameter combination as optimal using just the training set, and not the testing set. If the stochastic process we are evaluating has no memory, there is no reason to expect overfitting to induce negative performance because the association level between in-sample performance and out-of-sample performance is nearly zero. The problem of overfitting arises in the presence of memory, which is usually the case when working with financial time series (economic cycles, bubbles' bursts, structural breaks, etc).

If there is memory in the data, the more we optimize IS performance, the worse is OOS performance. It happens due to a compensation effect that generates a strongly negative relation between in-sample and out-of-sample performances. Because financial time series are known to exhibit memory, the consequence of overfitting is negative performance OOS (this can be easily shown running a Monte Carlo experiment on serially correlated returns).

Final Remarks

In this article, we highlighted common pitfalls that researchers and practitioners often make when they are backtesting investment strategies, and this is particularly useful for investors that use backtested performance to evaluate the performance of a portfolio. Even though backtests do have some drawbacks, we certainly cannot say that they are all useless. Backtests can be a very powerful tool if we stick to the math, take proper precautions in modelling and account for its limitations. Backtesting is indeed about the past and history does not repeat itself but remember that it often rhymes.

Hully Rolemberg

Quant Developer, MSc in Econometrics, and Economist. Previous experiences include teaching and economic research.

Gabriel Koike

Quant and Team Lead of the Quantitative Investment Strategies Team. Formerly, designed control laws algorithms as an engineer in the aerospace industry.

References

Bailey, David H., and Marcos Lopez de Prado. "The Sharpe ratio efficient frontier." Journal of Risk 15.2 (2012): 13.

Bailey, David H., et al. "Pseudomathematics and financial charlatanism: The effects of backtest over fitting on out-of-sample performance." Notices of the AMS 61.5 (2014): 458-471.

Bailey, David H., et al. "The probability of backtest overfitting." Journal of Computational Finance, forthcoming (2016).

Politis, Dimitris N., and Joseph P. Romano. "The stationary bootstrap." Journal of the American Statistical association 89.428 (1994): 1303-1313.

Disclosures

The opinions expressed in this publication are those of the authors. They do not purport to reflect the opinions or views of Daemon Investments or its members. The designations employed in this publication and the presentation therein do not imply the expression of any opinion whatsoever on the part of Daemon Investments.

This presentation ("Presentation") is furnished on a strictly confidential basis, prepared exclusively for the recipient to whom it was originally addressed, intended for informational purposes only and should not be considered an offer, solicitation, or recommendation to sell or buy any investment, security, or product. Any offer or solicitation will only be made by means of the applicable Fund documents, in jurisdictions in which such an offer would be lawful and only to investors who meet the suitability requirements. This product has not been created for investors who are U.S. Persons or any person acting on behalf of a U.S. Person.

Your acceptance of this Presentation constitutes your agreement to (i) maintain the confidentiality of its information and any information derived from it ("Confidential Information"); (ii) refrain from using the Confidential Information for any purpose other than monitoring your investment in the Fund; (iii) refrain from using the Confidential Information for purposes of trading any security, including securities of Daemon's portfolio; (iv) refrain from copying this Presentation without Daemon's prior consent; and, (v) promptly return this Presentation and any copies, or destroy any electronic copies, upon Daemon's request.

Past or simulated performance is not necessarily indicative of future results. Simulations and projections are statistical models based on data and assumptions to try to predict future scenarios that may affect the portfolio, therefore: (1) they are not free of errors; (2) scenarios are not guaranteed to occur; (3) they depend on data from external sources, not guaranteed by Daemon; (4) they do not guarantee returns or maximum loss exposure; and (5) they should not be used to support any administrative procedure before Supervisory or Regulatory Bodies.

Opinions and estimations provided herein may change at any time without prior notice. This Presentation is not intended to substitute for any investment, financial, legal, regulatory, accounting, tax or other professional advice, consultation, or service. Consult your tax advisor about the effect that taxes may have on potential returns. Daemon does not warrant or guarantee that Brazilian multi-market investment funds will be treated as long term investment funds by the Brazilian tax administration. From time to time, Daemon and/or any of its officers, directors or affiliates may have a long or short position in, or otherwise be directly or indirectly interested in, the securities, commodities, currencies, options, rights or warrants of companies or instruments mentioned herein. Daemon, its affiliated and related parties accept no liability whatsoever for the correctness, completeness or accuracy of the information contained herein or for any direct or consequential damages or loss arising from any use of the information contained in this document, although all information and opinions expressed herein have been obtained from sources believed to be reliable and in good faith.