TACTICAL INVESTMENT ALGORITHMS

Marcos López de Prado

This version: September 30, 2019

Marcos López de Prado is CIO of True Positive Technologies, LP, in New York, NY, and Professor of Practice at Cornell University's School of Engineering, in Ithaca, NY. E-mail: mldp@truepositive.com.

TACTICAL INVESTMENT ALGORITHMS

ABSTRACT

There are three fundamental ways of testing the validity of an investment algorithm against historical evidence: a) the walk-forward method; b) the resampling method; and c) the Monte Carlo method. By far the most common approach followed among academics and practitioners is the walk-forward method. Implicit in that choice is the assumption that a given investment algorithm should be deployed throughout all market regimes. We denote such assumption the "all-weather" hypothesis, and the algorithms based on that hypothesis "strategic investment algorithms" (or "investment strategies").

The all-weather hypothesis is not necessarily true, as demonstrated by the fact that many investment strategies have floundered in a zero-rate environment. This motivates the problem of identifying investment algorithms that are optimal for specific market regimes, denoted "tactical investment algorithms." This paper argues that backtesting against synthetic datasets should be the preferred approach for developing tactical investment algorithms. A new organizational structure for asset managers is proposed, as a tactical algorithmic factory, consistent with the Monte Carlo backtesting paradigm.

Keywords: Backtest overfitting, selection bias, multiple testing, quantitative investments, machine learning, all-weather hypothesis, strategic investment algorithm, tactical investment algorithm.

JEL Classification: G0, G1, G2, G15, G24, E44.
AMS Classification: 91G10, 91G60, 91G70, 62C, 60E.

Two major epistemological limitations prevent finance from becoming a science, at par with physics, chemistry or biology. First, finance does not comply with Popper's falsifiability criterion, because financial theories cannot be tested in a laboratory in controlled experiments. Claims such as "value and momentum factors explain the outperformance of stocks" cannot be proven wrong, even if they are. All researchers have is the outcome from a single realized path (a price time series) produced by an unknown data-generating process (DGP). We cannot draw millions of alternative paths from the same DGP and evaluate in how many instances value and momentum factors had explanatory power, while controlling for environmental conditions.

The second epistemological limitation afflicting finance is non-stationarity. Financial systems are extremely dynamic and complex, with conditions that quickly change over time. Financial cause-effect mechanisms are not invariant, due to changes in regulation, expectations, economic cycles, market regimes and other environmental variables. For instance, even if value and momentum factors truly explained the outperformance of stocks in the 20th century, that may no longer be the case as a result of recent technological, behavioral or policy changes. Perhaps value and momentum only worked under certain conditions that are no longer present. Consequently, claims made by financial economists are typically based on anecdotal information, and do not rise to the standard of scientific theories.

Due to these epistemological limitations, researchers rely on backtesting for developing investment algorithms. A backtest infers the performance of an investment algorithm under the general assumption that future observations will be drawn from the same DGP that produced past observations. In this paper, I explain the different types of backtesting methods, and the specific assumptions underlying each method. I also argue that one particular type of backtesting method can help address finance's epistemological limitations, and bring financial theories closer to scientific standards.

THE THREE TYPES OF BACKTESTS

In general terms, we can differentiate between three types of backtests. First, the walk-forward method (WF) assesses the performance of an investment algorithm under the assumption that history repeats itself *exactly*.[1] A first caveat of WF is that past time series merely reflect one possible path produced by the DGP. If we were to take a time machine, the stochastic nature of the DGP would produce a different path. Since WF backtests are not representative of the past DGP, there is no reason to believe that they are representative of the future DGP. Accordingly, WF is more likely to yield a descriptive (or anecdotal) than an inferential statement (see López de Prado [2018], chapter 11). A second caveat of WF is that the DGP is never stated: should the DGP change, the researcher will not be able to decommission the algorithm *before* it loses money, because she never understood the conditions that made the algorithm work.

The second type of backtest is the resampling method (RS), which addresses WF's first caveat. RS assesses the performance of an investment algorithm under the assumption that future paths can be simulated through the resampling of past observations. The resampling can be deterministic (e.g., jackknife, cross-validation) or random (e.g., subsampling, bootstrap).

[1] The main argument in favor of WF is that it prevents leakage from look-ahead information. However, if a walk-backwards backtest does not exhibit significantly better performance than a WF, look-ahead leakage is not a concern, making the main argument for WF rather weak.

Because RS can produce many different paths, where the historical is just one possibility, it allows us to consider more general scenarios consistent with the DGP. For instance, through a RS backtest we can bootstrap the distribution of the algorithm's Sharpe ratio, which is much more informative than the single-path Sharpe ratio derived by WF. Whereas it is trivial to overfit a WF backtest, it is more difficult to overfit a RS backtest. Still, resampling on a finite historical sample may not yield paths representative of the future (see López de Prado [2018], chapter 12).

The third type of backtest, the Monte Carlo method (MC), addresses both of WF's caveats. The MC method assesses the performance of an investment algorithm under the assumption that future paths can be simulated via Monte Carlo. MC requires a deeper knowledge of the DGP, derived from the statistical analysis of the observations or theory (e.g., market microstructure, institutional processes, economic links, etc.). For instance, economic theory may suggest that two variables are cointegrated, and empirical studies may indicate the range of values that characterize the cointegration vector. Accordingly, researchers can simulate millions of years of data, where the cointegration vector takes many different values within the estimated range. This is a much richer analysis than merely resampling observations from a finite (and likely unrepresentative) set of observations (see López de Prado [2018], chapter 13).

A PRACTICAL EXAMPLE OF MC BACKTEST

Consider a researcher that wishes to design a market making algorithm. Market microstructure theory tells us uninformed traders cause short-term mean reversion as a result of temporary market impact, and informed traders cause permanent impact on market prices. Informed traders arrive at the market at a rate μ and uninformed traders arrive at the market at a rate ε, where both rates can be modelled with a Poisson process. The statistical analysis of historical time series gives us a range of fluctuation of μ and ε, which can be used to simulate long series under various scenarios. For a given combination of μ and ε, MC allows us to derive the optimal market making algorithm, that is, the set of profit taking and stop loss levels that maximize the Sharpe ratio in a MC backtest. In contrast, WF and RS would backtest the overall performance of the market making algorithm, over all historical values of μ and ε, without allowing us to estimate the performance at specific pairs of μ and ε, and without allowing us to derive optimal market making algorithms for each specific pair.

Exhibit 1 shows the performance of a trading algorithm under various profit-taking and stop-loss scenarios, where the underlying price follows an Ornstein-Uhlenbeck process with a half-life of 5, zero drift and noise with unit variance (see López de Prado [2018], chapter 13). The half-life is so small that performance is maximized in a narrow range of combinations of small profit-taking with large stop-losses. In other words, the optimal trading rule is to hold an inventory long enough until a small profit arises, even at the risk of experiencing some 5-fold or 7-fold unrealized losses. Sharpe ratios are high, reaching levels of around 3.2. The worst possible trading rule in this setting would be to combine a short stop-loss with a large profit-taking threshold, a situation that market-makers avoid in practice. Performance is closest to neutral in the diagonal of the mesh, where profit-taking and stop-losses are symmetric.

[EXHIBIT 1 HERE]

Exhibit 2 shows what happens when the half-life increases from 5 to 10. The areas of highest and lowest performance spread over the mesh, while the Sharpe ratios decrease to levels around or below 2. This is because, as the half-life increases, so does the magnitude of the autoregressive coefficient, thus bringing the process closer to a random walk. For a sufficiently long half-life, even the optimal combination of profit-taking and stop-loss levels yield an unacceptably low return on risk.

[EXHIBIT 2 HERE]

FOUR UNIQUE ADVANTAGES OF MC BACKTESTS

MC offers four critical advantages over WF and RS. First, MC backtests help address the first epistemological limitation of finance, because they allow researchers to conduct randomized controlled experiments. Admittedly, these experiments require the assumption of a particular DGP, but at least that DGP is explicitly stated (unlike in the WF backtests published in financial journals). In MC backtests, the researcher declares the hypothesis underlying her findings. If the investor believes that the true DGP is different, she just needs to propose an alternative DGP and repeat the analysis. We can consider MC backtests a particular case of *Ersatz tests*, where statistical methods are tested on computer-generated data from known models (Jarvis et al. [2017]).

Second, MC backtests help address the second epistemological limitation of finance, because the researcher does not need to assume that the DGP is immutable. Instead, the discovery is connected to a particular DGP, where realizations may be drawn from different DGPs over time. In other words, MC backtests allow us to develop "tactical investment algorithms," as opposed to the "strategic investment algorithms" developed with the help of WF or RS.[2] The probability that a particular DGP is producing the realizations can be evaluated statistically, which allows researchers to commission or decommission tactical algorithms as conditions evolve.

Third, MC backtests enable the incorporation of priors, which inject information beyond what we could have learned from a finite set of observations. When these priors are motivated by economic theory, MC offers a powerful tool to simulate the most likely scenarios, even if some of those scenarios have not been observed in the past. Unlike WF or RS, MC backtests can help us develop tactical algorithms to be deployed in the presence of black swans.

Fourth, the length of MC backtests can be expanded for as long as needed to achieve a targeted degree of confidence. This is helpful in that MC backtests avoid the indetermination inherent to working with finite datasets.

THE CRITICISM OF MC BACKTESTS

Investors are sometimes skeptical of MC backtests, because they compute the performance of investment algorithms on synthetic data, which may not be representative of future realizations of the true DGP. This skepticism is misplaced, for two reasons: (a) estimating a DGP is not

[2] In recent years, it has been proven fashionable for some asset managers to promote certain investment factors through long WF backtests (in some cases, covering over a hundred years). Consider the validity of that work when, for instance, the current environment of negative interest rates has never been experienced before. In contrast, it is straightforward to conduct a MC backtest on data simulated by a DGP with negative interest rates.

necessarily a harder problem than forecasting the markets. It is intellectually incoherent to assume that, on one hand, statistical methods can lead to successful investment outcomes but, on the other hand, statistical methods cannot identify a DGP; (b) the observations used by WF and RS are unlikely to reoccur in the future exactly as simulated, and the paths generated by MC are not necessarily less likely.

Another concern is that a researcher may select a DGP that is particularly favorable to the investment algorithm. This concern is also misplaced: the MC method explicitly declares the assumptions underlying the performance simulations, so if the DGP is unrealistically favorable to the algorithm, the investor can object. In contrast, the WF and RS methods imply those assumptions through the selection of the historical dataset used by the simulations, obfuscating the dangers of selection bias and confirmation bias.

EXAMPLES OF DGPs

A Monte Carlo randomly samples new (unobserved) datasets from an estimated population or DGP, rather than from an observed dataset (like a bootstrap would do). Monte Carlo experiments can be parametric or non-parametric. An instance of a parametric Monte Carlo is a regime-switching time series model (Hamilton [1994]), where samples are drawn from alternative processes, $n = 1, \ldots, N$, and where the probability $p_{t,n}$ of drawing from process n at time t is a function of the process from which the previous observation was drawn (a Markov chain). Expectation-maximization algorithms can be used to estimate the probability of transitioning from one process to another at time t (the transition probability matrix). This parametric approach allows researchers to match the statistical properties of the observed dataset, which are then replicated in the unobserved dataset (see Franco-Pedroso et al. [2019]).

One potential caveat of parametric Monte Carlo is that the DGP may be more complex than a finite set of algebraic functions can replicate. When that is the case, non-parametric Monte Carlo experiments may be of help, through the use of variational autoencoders, self-organizing maps, or generative adversarial networks (De Meer Pardo [2019]). These methods can be understood as non-parametric, non-linear estimators of latent variables (similar to a non-linear PCA). An autoencoder is a neural network that learns how to represent high-dimensional observations in a low-dimensional space. Variational autoencoders have an additional property which makes their latent spaces continuous. This allows for successful random sampling and interpolation and, in turn, their use as a generative model. Once a variational autoencoder has learned the fundamental structure of the data, it can generate new observations that resemble the statistical properties of the original sample, within a given dispersion (hence the notion of "variational"). A self-organizing map differs from autoencoders in that it applies competitive learning (rather than error-correction), and it uses a neighborhood function to preserve the topological properties of the input space. Generative adversarial networks train two competing neural networks, where one network (called a generator) is tasked with generating simulated observations from a distribution function, and the other network (called a discriminator) is tasked with predicting the probability that the simulated observations are false given the true observed data. The two neural networks compete with each other, until they converge to an equilibrium. The original sample on which the non-parametric Monte Carlo is trained must be representative enough to learn the general characteristics of the DGP, otherwise a parametric Monte Carlo approach should be preferred. See López de Prado [2019] for additional details.

THE TACTICAL ALGORITHMIC FACTORY

The WF and RS backtesting methods attempt to find "all-weather" algorithms, that is, strategic investment algorithms that are not associated with a particular DGP, and are deployed under all market conditions. The notion of strategic (all-weather) investment algorithms is inconsistent with the fact that markets go through regimes, during which some algorithms are expected to work and others expected to fail. Given that markets are adaptive and investors learn from mistakes, the likelihood that truly all-weather algorithms exist is rather slim (an argument often wielded by discretionary portfolio managers). And even if all-weather algorithms existed, they are likely to be a rather insignificant subset of the population of algorithms that work across one or more regimes.

In contrast to WF and RS backtests, MC backtests help us define the precise sensitivity of an investment algorithm to the characteristics of each DGP. Once we understand what characteristics make the algorithm work, we can deploy it tactically, while monitoring the idoneity of market conditions, and derive the appropriate ex-ante risk allocations. When used in this way, MC backtests allow us to *trade the algorithms rather than the markets*. Under this investment paradigm, a firm will develop as many tactical investment algorithms as possible (López de Prado [2018], chapter 1), and then deploy only those algorithms that are certified to work under the prevalent market conditions. These algorithms are DGP-specific, not instrument specific: the same algorithm will be deployed tactically on different instruments over time, when those instruments temporarily follow the DGP associated with that algorithm. The main difference between the tactical algorithmic factory (TAF) approach and the strategic algorithmic factory (SAF) approach is that TAF's objective is to develop DGP-specific algorithms, which are not required to work all the time. Instead, TAF's algorithms only need to work during the DGP for which they have been certified.

DGP IDENTIFICATION

MC backtests allow researchers to pose the algorithm selection problem in terms of a DGP identification problem. This is advantageous, because finding an algorithm that works well across all possible DGPs is much more challenging than estimating what is the current DGP (which in turn determines the algorithm that should be run at a given point in time). Also, from a mathematical perspective, identifying the optimal algorithm associated with a particular DGP is a well-defined problem.[3]

One practical way of identifying the prevailing DGP is as follows: First, through MC backtests, develop many tactical investment algorithms for a wide range of DGPs. Second, select a sample of recent market performance. Third, evaluate the probability that the sample of recent market performance was drawn from each of the studied DGPs. This probability can be estimated through different methods, such as the Kolmogorov-Smirnov test, the Wasserstein distance, or the Kullback-Leibler divergence. The resulting distribution of probability can then be used to allocate risk across the algorithms developed by the TAF. In other words, an ensemble of optimal strategies is deployed, and not only the most likely optimal strategy.

[3] Most journal articles promote investment algorithms without stating the DGP that those algorithms supposedly exploit. Without knowing the DGP, we cannot know the conditions under which that algorithm is supposed to be run, or when to decommission it.

In practice, it takes only a few recent observations for the estimated distribution of probability to narrow down the likely DGPs. The reason is, we are comparing two samples, where the synthetic one is comprised of potentially millions of datapoints, and it typically does not take many observations to discard what DGPs are inconsistent with recent observations.

Another possibility is to create a basket of securities with a returns distribution that matches the distribution of a given DGP. Under this alternative implementation, rather than estimating the probability that a security follows a DGP, we create a synthetic security (as a basket of securities) for which a given algorithm is optimal.

One virtue of running an ensemble of optimal algorithms is that the ensemble strategy does not correspond to any particular DGP. This allows the ensemble strategy to dynamically and smoothly transition from one DGP to another, and even profit from a never-seen-before DGP.

CONCLUSIONS

In this paper I have argued that MC backtests offer to financial researchers the possibility of conducting randomized controlled experiments. Absent financial laboratories, this is as close as finance can get to the Popperian criterion of falsifiability.

An MC backtest can be understood as a certification of the performance of an algorithm subject to certain declared environmental conditions, similar to how an engineer would certify the performance of a type of equipment. In contrast with the WF and RS methods, MC backtests inform us about the conditions under which the tactical investment algorithm should be deployed. This information also helps investors pinpoint the circumstances under which the algorithm is most vulnerable, when the algorithm should be decommissioned, and how much risk should be allocated to it.

Given that markets are adaptive and investors learn from mistakes, the likelihood that truly all-weather algorithms exist is rather slim (an argument often wielded by discretionary portfolio managers). And even if all-weather algorithms existed, they are likely to be a rather insignificant subset of the population of algorithms that work across one or more regimes. Accordingly, asset managers should embrace the TAF paradigm, hence developing as many tactical investment algorithms as possible, through MC backtesting.

REFERENCES

De Meer Pardo, F. (2019): "Enriching Financial Datasets with Generative Adversarial Networks." Working paper. Available at http://resolver.tudelft.nl/uuid:51d69925-fb7b-4e82-9ba6-f8295f96705c

Franco-Pedroso, J., J. Gonzalez-Rodriguez, J. Cubero, M. Planas, R. Cobo, and F. Pablos (2019): "Generating Virtual Scenarios of Multivariate Financial Data for Quantitative Trading Applications." *The Journal of Financial Data Science*, 1(2), pp. 55-77. Available at https://doi.org/10.3905/jfds.2019.1.003

Hamilton, J. (1994): *Time Series Analysis*. Princeton University Press, First edition.

Jarvis, S., J. Sharpe, and A. Smith (2017): "Ersatz Model Tests." *British Actuarial Journal*, 22(3), pp. 490-521.

López de Prado, M. (2018): *Advances in Financial Machine Learning*. First edition, Wiley. https://www.amazon.com/dp/1119482089

López de Prado, M. (2019): *Systems and Methods for a Factory that produces Tactical Investment Algorithms through Monte Carlo Backtesting*. United States Patent and Trademark Office, Application No. 62/899,164.

López de Prado, M. (2019): *Machine Learning for Asset Managers*. First edition, Cambridge. Forthcoming.

EXHIBITS

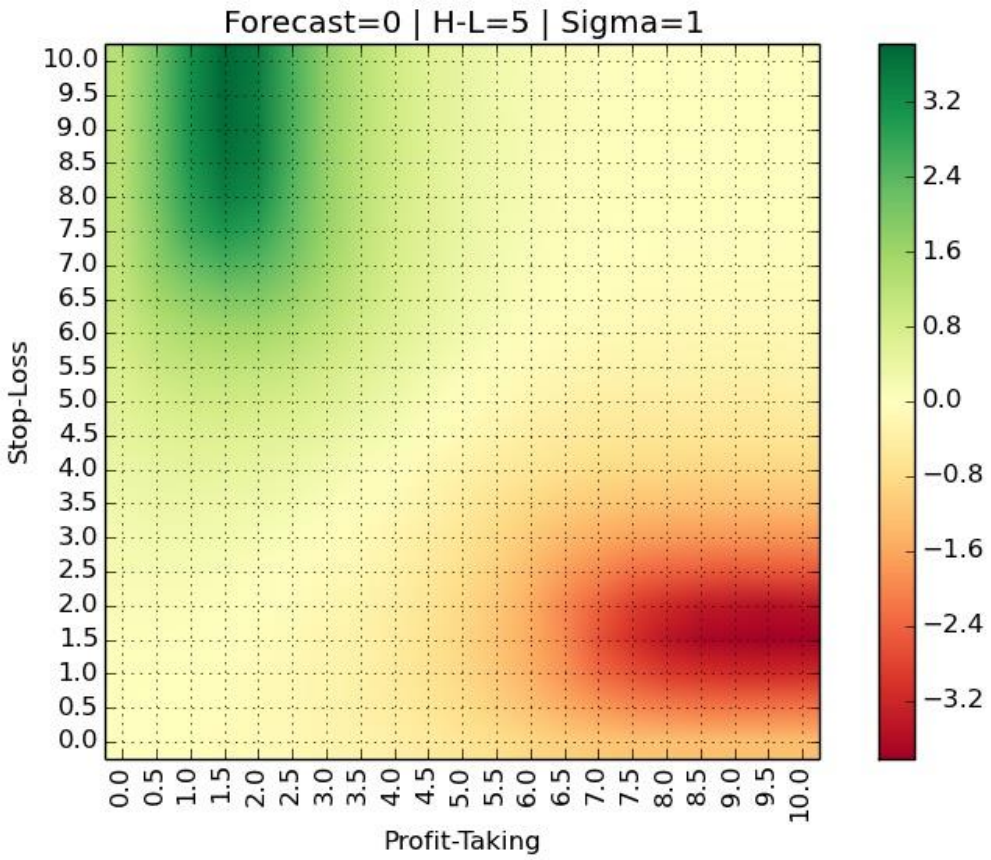

Exhibit 1 – Sharpe ratios associated with various combinations of profit-taking and stop-loss, for an Ornstein-Uhlenbeck process with half-life 5, no drift, and noise with unit variance

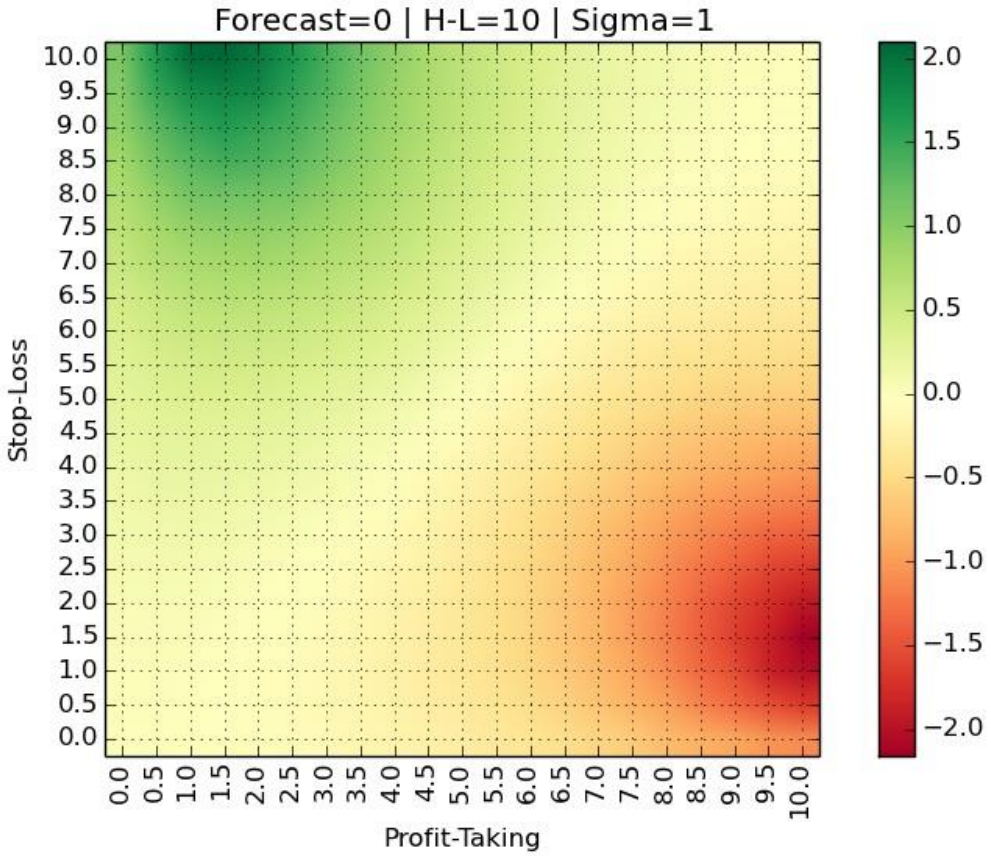

Exhibit 2 – Sharpe ratios associated with various combinations of profit-taking and stop-loss, for an Ornstein-Uhlenbeck process with half-life 10, no drift, and noise with unit variance