Backtest Overfitting on Out-of-Sample Performance

Ting-Yen Tsai

June 5th, 2020

1 Introduction

Nowadays, people leverage their trading ideas, strong computing ability and machine learning techniques to find new investment strategy and optimal capital allocation. When evaluating the quality of an investment strategy, academia and industry professionals often use a *backtest*, which is a historical simulation that calculate the return the strategy would have generated over a certain time period.
As the trained model may capture and overfit particular observations rather than the whole structure, so it would make really bad predictions for new datasets. One simple idea to avoid overfitting is that we use a part of data as *training data* to train our trading strategy, and use other *test data* to if our strategy really have the ability to capture growth. Generally, if the strategy has consistent *in-sample* and *out-of-sample* performances, then people believe that it is a good strategy that can work for other time period and even in the future. However, Bailey, Borwein, Lopez de Prado and Zhu mentioned some issues in the implementation of machine learning in their paper, *Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample performance* [1].
In this paper, we will discuss backtest overfitting and its negative effect on out-of-sample performance. Since this is a topic related to statistic and finance, we will start from introducing the basic methodology in machine learning as well as some terms in finance and time series analysis, and then explore backtest overfitting and end with a practical simulation done by an online platform. We assume the reader to have background knowledge about inferential statistics and probability.

1.1 Machine Learning Basics

Machine learning is the study of automatically improving computer algorithm by accumulating experience from currently available data, and its ultimate goal is to achieve highly accurate predictions for unknown responses [2]. There are many machine learning techniques. Two main streams are supervised learning and unsupervised learning. In supervised learning, we collect a set of training sample $(X_1, Y_1), ..., (X_n, Y_n)$ that is independent and identically distributed with the same distribution of (X, Y), where X is a vector containing explanatory variables and Y is a scalar response. Our goal is to find a prediction rule $f(X) = E[Y|X]$. Common approaches include (generalized) linear regression, regression tree, K-Nearest Neighbor. In unsupervised learning, we do not have assumption about the

causal relations between each variables and reduce human supervision to the minimum. Principle component analysis (PCA), cluster analysis, K-means are some popular unsupervised learning methods. The practitioners in finance industry primarily use supervised learning, because it is widely believed that there are some key parameters that can affect the price of assets based on economic theory. Thus, we would restrict our discussion to supervised learning in this paper.
The following is a general procedure for machine learning:

1. Split the all data into two sets: training set and test set. A rule of thumb for the ratio is 1:4.
2. Train the model by the training set, and calculate the *in-sample* mean squared error. For example, if we are training a simple linear regression model, we get the coefficients for each parameter.
3. Predict the result for the test set with the trained model, and calculate the *out-of-sample* mean sqaured error (MSE). If the *in-sample* and *out-of-sample* MSE are consistent, then the model is good.
4. Re-split the data to train set and test set randomly and compare the average *in-sample* and *out-of-sample* MSE to test the model's robustness. This step is often called *cross-validation*.

1.2 Finance Terms

In this subsection, we introduce two common performance measures for investment strategy.

Definition 1 (Ex Ante Sharpe Ratio [3])**.** *We define the Ex Ante Sharpe Ratio as*

$$SR = \frac{\widetilde{R_p} - \widetilde{R_f}}{\sigma_d},$$

where $\widetilde{R_p}$ is the expected return of the fund, $\widetilde{R_f}$ is the expected return of risk-free asset. $\widetilde{R_p} - \widetilde{R_f}$ defines differential return d and σ_d is the predicted standard deviation of d.

Definition 2 (Information Ratio [4])**.** *We define the Intermation Ratio as*

$$IR = \frac{\widetilde{R_p} - \widetilde{R_b}}{\sigma_e},$$

where $\widetilde{R_p}$ is the expected return of the portfolio, $\widetilde{R_b}$ is the expected return of a benchmark. $\widetilde{R_p} - \widetilde{R_b}$ defines excess return e and σ_e is the predicted standard deviation of e.

One remark here is Sharpe Ratio use the expected return of risk-free asset (e.g. 3-month U.S. Treasury Bill), while the Information Ratio use the expected return of a benchmark (e.g. S&P 500).

1.3 Autoregressive Process

In this subsection, we introduce the definition of an autoregressive process.

Definition 3. *A stochastic process is a family of random variables* $\{X_t\}$.

Definition 4. *If the stochastic process* $\{X_t\}$ *has the mean function* $\mu(t) = E[X_t]$, *then its autocovariance of any* t_1 *and* t_2 *is given by* $\gamma(t_1, t_2) = Cov[X_{t_1}, X_{t_2}]$.

Definition 5. *A stochastic process is weakly stationary if the mean function* $\mu(t) = \mu$ *for all* t *and the autocovariance* $\gamma(t_1, t_2) = \gamma(t_2 - t_1, 0)$ *for any* t_1, t_2.

Definition 6. *An autoregressive process* $\{X_t\}$ *of order* p *is given by*

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + ... + \phi_p X_{t-p} + Z_t,$$

where Z_t *is the white noise that i.i.d. to* $N(0, \sigma^2)$.

Theorem 1. *A first-order autoregression process is weakly stationary if* $E[X_t] = 0$ *and* $|\phi_1| < 1$.

Proof.

$$\begin{aligned}
X_t &= \phi_1 X_{t-1} + Z_t = \phi_1(\phi_1 X_{t-2} + Z_{t-1}) + Z_t \\
&= \phi_1^t X_0 + \sum_{i=0}^{t-1} \phi_1^i Z_{t-i} \\
E[X_t] &= \phi_1^t E[X_0] + \sum_{i=0}^{t-1} \phi_1^i E[Z_{t-i}] = \phi_1^t E[X_0] \\
E[X_t] &= \phi_1^t E[X_0] = 0 \\
X_{t+h} &= \phi_1^h X_t + \sum_{i=1}^{h-1} \phi_1^i Z_{t+h-i} \\
\gamma(t, t+h) &= Cov[X_t, X_{t+h}] = Cov(X_t,\ \phi_1^h X_t + \sum_{i=1}^{h-1} \phi_1^i Z_{t+h-i}) \\
&= \phi_1^h Var(X_t) = \frac{\phi_1^h \sigma^2}{1 - \phi_1^2}
\end{aligned}$$

For any first-order autoregression process, $E[X_t] = 0$ for all t, and its autocovariance converges if and only if $|\phi_1| < 1$ by geometric series. It proves the theorem. □

2 Backtest Overfitting

We assume that a strategy's differential redurn d is independent and identically distributed as

$$d \sim N(\mu, \sigma^2), \tag{1}$$

where N represents a Normal distribution with mean μ and variance σ^2.
Then we can represent *annualized* Sharpe Ratio as

$$SR = \frac{\mu}{\sigma} \cdot \sqrt{q}, \tag{2}$$

where q is the number of returns per year.
However, since μ and σ are unknown, we have to estimate the Sharpe Ratio as $\hat{SR} = \frac{\hat{\mu}}{\hat{\sigma}} \cdot \sqrt{q}$, where $\hat{\mu}$ and $\hat{\sigma}$ are sample mean and sample standard deviation. If we apply backtest to the same data for many times, we are likely to get a false positive and believe that strategy has a high Sharpe Ratio not only *in-sample* but also *out-of-sample* (not overfitting). For instance, we would like to find a strategy that will generate a SR bigger than 2. Suppose we come up with twenty candidate strategies, use Akaike Information Criterion, a common measure for goodness of fit, to check for overfitting and set our confidence level at 5%, then we may expect one out of twenty trials will achieve our goal, yet this success may be pure coincidence [1]. Since we do not have the information about the number of trials behind a "successful" investment strategy, we may neglect the possibility, or even the fact of model overfitting.

In order to formalize and quantify backtest overfitting, we define it as the following.

Definition 7 (Backtest Overfitting [5])**.** *We say that the backtest strategy selection process overfits if a strategy with optimal performance IS has an expected ranking below the median OOS. By the Bayes' Theorem,*

$$\sum_{n=1}^{N} E[\bar{r_n} | r \in \Omega_n^*] Prob[r \in \Omega_n^*] \leq N/2,$$

where r is a vector representing rank IS, $\bar{r}$ is a vector representing rank OOS, $\Omega_n^* = \{f \in \Omega | f_n = N\}$ is a subset of the space includes all permutation of $(1, 2, ..., N)$ indicating the rank of the strategy, and N denotes the size of each random sample.

Now we want to show that as people do more trial for the same model, the maximum of the expected return will increase, which can be misleading because we do not change anything about the model.

Theorem 2. *Given a sample of i.i.d. random variables, $x_n \sim Z, n = 1, ..., N$, where Z is the cumulative density function of the Standard Normal Distribution, the expected maximum of the sample, $E[max_N] = E[max\{x_N\}]$, can be approximated for a large N as*

$$E[\max_N] \approx (1 - \gamma) Z^{-1}[1 - \frac{1}{N}] + \gamma Z^{-1}[1 - \frac{1}{N} e^{-1}],$$

where $\gamma \approx 0.5772156649...$ is the Euler-Mascheroni constant and $N \gg 1$.

Proof. Since x_n is independent and identically distributed and $x_n \sim Z$, we apply the Fisher-Tippett-Gnedenko Theorem. If a sequence of pairs of real number (a_n, b_n) exists s.t. each $a_n > 0$ and

$$\lim_{n\to\infty} \text{Prob}(\frac{\max_n - b_n}{a_n} \leq x) = F(x),$$

where F is a non-degenerate distribution function, then the limit distribution F belongs to either the Gumbel, the Frechet or the Weibull family. For Standard Normal Distribution, we have $F(x) = e^{-e^{-x}}$ (the cumulative density function of the Standard Gumbel Distribution), and $a_n = Z^{-1}[1-\frac{1}{n}e^{-1}]-b_n > 0$ and $b_n = Z^{-1}[1-\frac{1}{n}]$, where Z^{-1} is the inverse of Z [6] [7]. Also, from [6] we have

$$\lim_{n\to\infty} E[\frac{\max_n - b_n}{a_n}] = \gamma,$$

where γ is the Euler-Mascheroni constant. Thus, for N sufficiently large ($N \gg 1$), we have

$$E[\max_N] \approx b_N + \gamma a_N = (1-\gamma)Z^{-1}[1-\frac{1}{N}] + \gamma Z^{-1}[1-\frac{1}{N}e^{-1}].$$

□

From this theorem, we can see as the number of trial N increases, $E[\max_N]$ will increase and lead to a high Sharpe Ratio if we only pick out the sample that generate the highest return. We can apply this theorem to the condition when we observe a longer period y. From the previous result in (2), now $\hat{SR} \to N(0, \sqrt{y})$. By rescaling the standard deviation, we have

$$E[\max_N] \approx \frac{1}{\sqrt{y}}\{(1-\gamma)Z^{-1}[1-\frac{1}{N}] + \gamma Z^{-1}[1-\frac{1}{N}e^{-1}]\}, \tag{3}$$

where $\gamma \approx 0.5772156649...$ is the Euler-Mascheroni constant and $N \gg 1$.
By solving y in (3), we now define Minimum Backtest Length (MinBTL) as the following.

Theorem 3. *The Minimum Backtest Length (MinBTL, measured in years) needed to avoid selecting a strategy with an IS Sharpe Ratio of* $E[\max_N]$ *among* N *independent strategies with an expected OOS Sharpe Ratio of zero is*

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

3 Relationship Between *In-Sample* and *Out-of-Sample* Performance

So far we show that practitioners may intentionally select strategies that pass the AIC test by chance. Now would like to take a step forward the show that there are indeed a *negative* relation between *in-sample (IS)* and *out-of-sample (OOS)* performances. We will first consider a special case that does not cause OOS worse even though we still overfit, then show that there is a tradeoff between IS and OOS performances under global constraint and serial dependence.

3.1 Random Walk Model

We first generate N Gaussian random walks whose size is T, and each path m_τ can be obtained as a cumulative sum of Gaussian draws:

$$\Delta m_\tau = \mu + \sigma\epsilon_\tau, \tag{4}$$

where μ is the *drift term* that decide the trend of the process, and $\epsilon_\tau \sim N(0,1), \tau = 1, ..., T$ are the *random shocks* on every single step.
In [1], the authors assumed $\mu = 0, \sigma = 1$ and $T = 1000$, covering a period of one year with about four observation per trading day. We divide these paths into two disjoint paths into two disjoint samples of equal size 500, and use the first 500 samples as training set (*in-sample*) and the others as test set (*out-of-sample*). The result shows that even though the model overfits IS, the performance of OOS has nothing to do with that of IS, so there are no clear evidence to show negative effects.

3.2 Global Constraint

However, if we rerun the same Monte Carlo experiment as before, and center each performance path m_τ to match a mean μ (remove one degree of freedom), by defining the recenter variables $\overline{\Delta m_\tau}$:

$$\overline{m_\tau} = \Delta m_\tau + \mu - \frac{1}{T}\sum_{\tau=1}^{T} \Delta m_\tau \tag{5}$$

Theorem 4. *Given two different random samples (A and B) of the same model, where* $\sigma^A_{IS} = \sigma^A_{OOS} = \sigma^B_{IS} = \sigma^B_{OOS}$ *imposing a global constraint* $\mu_A = \mu_B$*, implies that*

$$SR^A_{IS} > SR^B_{IS} \Longleftrightarrow SR^A_{OOS} < SR^B_{OOS}$$

Proof. Suppose A and B are two random samples of the same process $\{\Delta m_\tau\}$, and they have same size, means and standard deviations. We take δ of each sample to be the training set (*in-sample*), and the remainder to be the test set (*out-of-sample*). Obviously, μ_A should lie between μ^A_{IS} and μ^A_{OOS}, and μ_B should lie between μ^B_{IS} and μ^B_{OOS}. From the global constraint $\mu_A = \mu_B$, we have

$$\mu_A = \delta\mu^A_{IS} + (1-\delta)\mu^A_{OOS} = \mu_B = \delta\mu^B_{IS} + (1-\delta)\mu^B_{OOS}$$

$$\mu^A_{IS} - \mu^B_{IS} = \frac{1-\delta}{\delta}(\mu^B_{OOS} - \mu^A_{OOS})$$

Then if $\mu^A_{IS} > \mu^B_{IS}$, then $\mu^A_{OOS} < \mu^B_{OOS}$. By dividing $\sigma = \sigma^A_{IS} = \sigma^A_{OOS} = \sigma^B_{IS} = \sigma^B_{OOS}$, we have

$$SR^A_{IS} > SR^B_{IS} \Longleftrightarrow SR^A_{OOS} < SR^B_{OOS}$$

□

From Theorem 4, we know that for a model, the higher the *in-sample* Sharpe Ratio, the lower the *out-of-sample* Sharpe Ratio. This result is massive because it shows that calibrating the parameters to optimize IS performance will lead to a bad OOS performance.

3.3 Serial Dependence

Let's assume that we have a prior belief: the price of a financial asset today are related to not only the trend μ, but also the price yesterday. Now we abandon the random walk model the previous subsections, and model the performance path $\{m_\tau\}$ as a first-order autoregressive process as the following.

$$m_\tau = (1-\phi)\mu + \phi m_{\tau-1} + \sigma\epsilon_\tau \tag{6}$$

$$\Delta m_\tau = (1-\phi)\mu + (\phi-1)m_{\tau-1} + \sigma\epsilon_\tau, \tag{7}$$

where $\phi \in (0,1)$ is the autoregressive coefficient that determines the contribution of the previous term $_{\tau-1}$ to the very term m_τ, μ is the *drift term* that indicates the trend of the process, and $\epsilon_\tau \sim N(0,1), \tau = 1, ..., T$ are the *random shocks* for every single step.
In order to draw conclusion for this case, we have to observe enough number of steps, because at the beginning of the process, there are periods where the process deviates from the expected value. Thus, we introduce a concept called *half-life*, which is the number of observations τ such that $E[m_\tau] = \frac{\mu+m_0}{2}$, where μ is the mean and m_0 is the initial value of the process.

Theorem 5. *The half-life period of a first-order autoregressive process with autoregressive coefficient $\phi \in (0,1)$ occurs at*

$$\tau = -\frac{\ln 2}{\ln \phi}.$$

Proof. Since the random shocks $\epsilon_\tau \sim N(0,1)$ and is independently and identically distributed,

$$\begin{aligned} E[m_\tau] &= (1-\phi)\mu + \phi m_{\tau-1} + 0 \\ &= (1-\phi)\mu + \phi[(1-\phi)\mu + \phi m_{\tau-2}] \\ &= (1-\phi)(1+\phi+\phi^2+...+\phi^{\tau-1})\mu + \phi^\tau m_0 \\ &= (1-\phi^\tau)\mu + \phi^\tau m_0 \end{aligned}$$

By plugging the above expression to the definition of half life,

$$\begin{aligned} E[m_\tau] = \frac{\mu+m_0}{2} &= (1-\phi^\tau)\mu + \phi^\tau m_0 \\ -\mu + m_0 &= 2(-\mu+m_0)\phi^\tau \\ \phi^\tau = \frac{1}{2} &\to \tau = -\frac{\ln 2}{\ln \phi} \end{aligned}$$

□

Theorem 6. *Given two different random samples (A and B) of the same model, where* $\sigma^A_{IS} = \sigma^A_{OOS} = \sigma^B_{IS} = \sigma^B_{OOS}$*, and the performance series follows the same first-order autoregressive stationary process,*

$$SR^A_{IS} > SR^B_{IS} \Longleftrightarrow SR^A_{OOS} < SR^B_{OOS}.$$

Proof. Suppose A and B are two random samples of a first-order autoregressive process and split A, B to two sets (training set and test set). The training set is the first δT steps in the process, and the test set is the remainder. From the proof of Theorem 5, we obtain

$$E[m_T] = (1 - \phi^T)\mu + \phi^T m_{\delta T}$$

$$E[m_T] - m_{\delta T} = (1 - \phi^T)(\mu - m_{\delta T})$$

Since $\phi \in (0, 1)$, $1 - \phi^T > 0$ and $\sigma^A_{IS} = \sigma^B_{IS}$, $SR^A_{IS} > SR^B_{IS} \Longrightarrow m^A_{\delta T} > m^B_{\delta T}$. Since $\mu_A = \mu_B$, $m^A_{\delta T} > m^B_{\delta T} \Longrightarrow E[m^A_T] - m^A_{\delta T} < E_[m^B_T] - m^B_{\delta T}$. By $\sigma^B_{IS} = \sigma^B_{OOS}$, we have

$$SR^A_{IS} > SR^B_{IS} \Longleftrightarrow SR^A_{OOS} < SR^B_{OOS}.$$

□

4 A Practical Simulation

In this section, we provide a detailed implementation about the procedure proposed in [1] that give more evidences to our previous theoretical results.

4.1 Algorithm

1. Generate a dataset either by drawing N samples with size T from Standard Normal Distribution or giving real historical market data.

2. Add a number of parameters to the model. In this case, we add *Entry Day* (the business day of the month when we enter a position), *Holding Period* (the number of days that we hold the position), *Stop Loss* (we will exit the position when the loss has reached this trigger) and *Side* (long or short).

3. Find the optimal value for each parameter that gives the highest Sharpe Ratio based on the given training data.

4. Execute trained strategy on *out-of-sample* data and calculate the Sharpe Ratio.

5. Visualize the relationship between *in-sample* and *out-of-sample* Sharpe Ratio.

4.2 Result

We take advantage of the online interactive program at http://datagrid.lbl.gov/backtest/index.php#run, and discuss the preformance of the following results in hindsight.

Example 1. The table below is the setting and result of the first example.

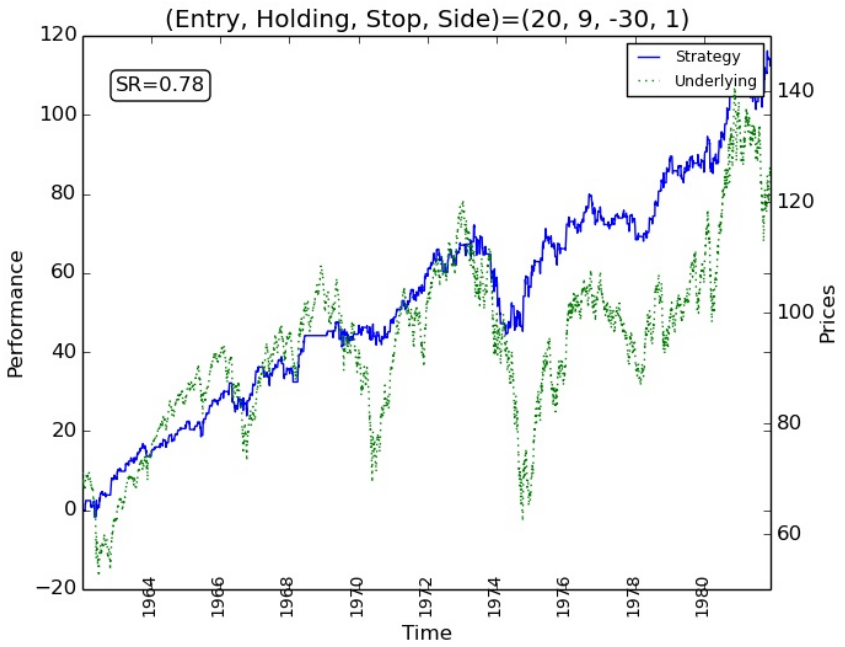
Setting		**Result**	
Maximum Holding Period	15	Entry Date	20
Maximum Stop Loss	30 (%)	Holding Period	9
Sample Length	5000	Stop Loss	30 (%)
Standard Deviation	N/A	Side	1
Seed	N/A	IR Sharpe Ratio	0.78
Real-World Data	Yes	OOS Sharpe Ratio	0.23

As we can see from Figure 1, while SR_{IS} is 0.78, SR_{OOS} is merely 0.23. *IS* is much more volatile at the beginning than *OOS*. This "long" strategy follows the price well at the beginning, but it is not immune to the price crash later on because it sets its *Stop Loss* at 30%.

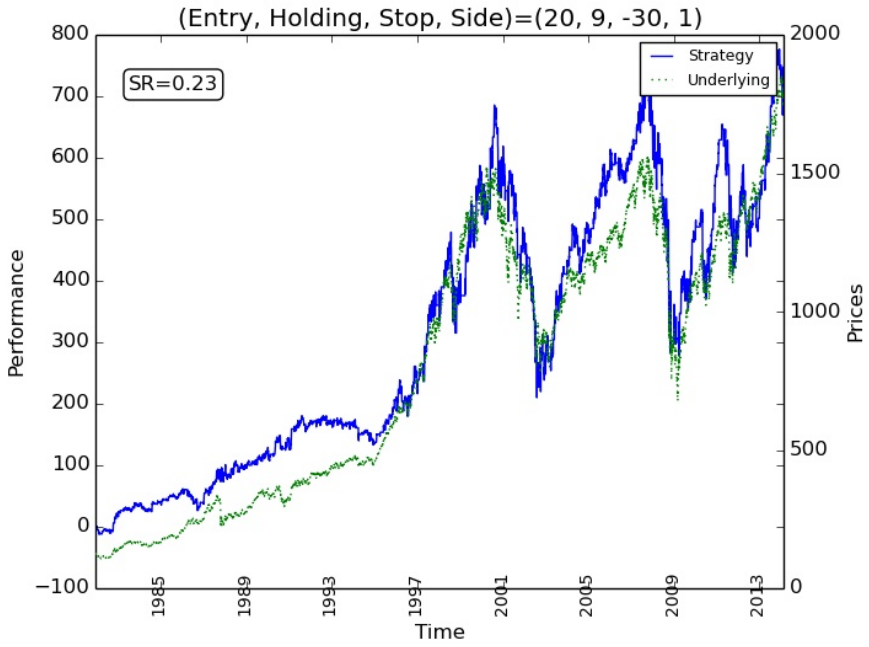
Example 2. The table below is the setting and result of the second example.

Setting		**Result**	
Maximum Holding Period	8	Entry Date	5
Maximum Stop Loss	20 (%)	Holding Period	8
Sample Length	1000	Stop Loss	3 (%)
Standard Deviation	1	Side	-1
Seed	10	IR Sharpe Ratio	1.91
Real-World Data	No	OOS Sharpe Ratio	-0.21

As we can see from Figure 2, while SR_{IS} is 1.91, SR_{OOS} is -0.21. The behavior of the underlying asset is really different between *IS* and *OOS*. The "short" strategy sucessfully captures high profit by correctly predict the direction *IS*, but since there are more volatile *OOS* and we get a negative Sharpe Ratio and strict *Stop Loss* trigger, so we may suspect that setting *Entry* at 5 is not optimal.

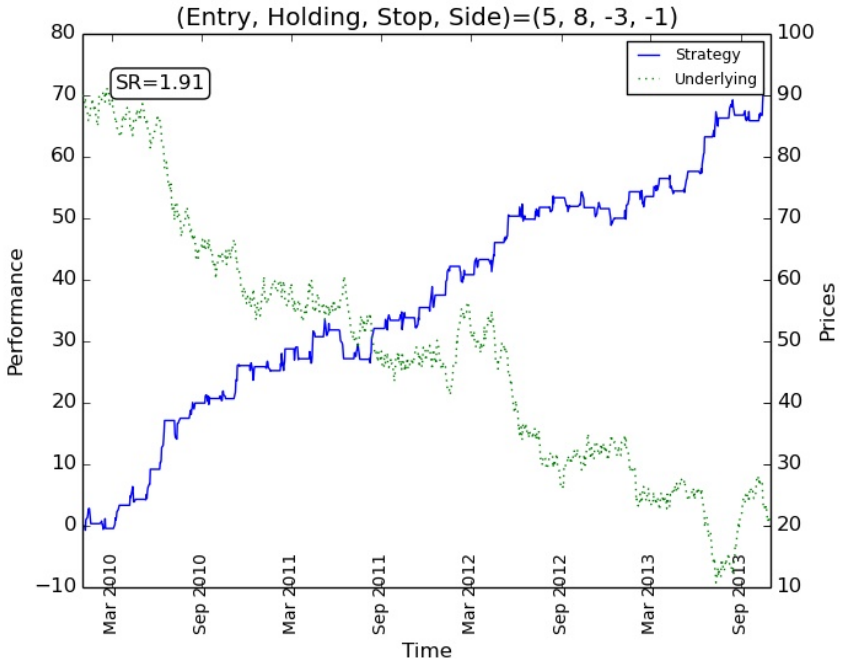


(a) IS Performance

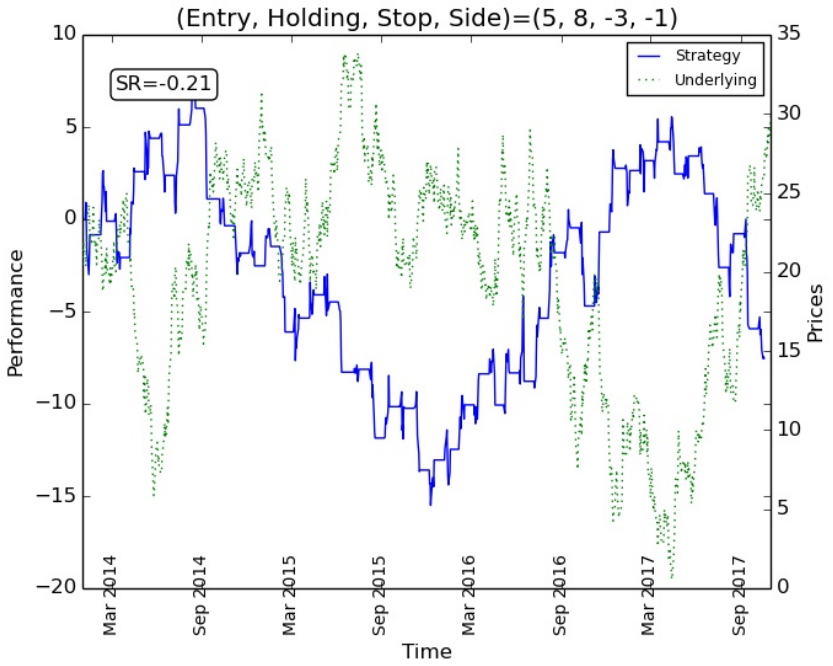


(b) OOS Performance

Figure 1: Example 1



(a) IS Performance



(b) OOS Performance

Figure 2: Example 2

5 Conclusion

Backtest overfitting is detrimental to developing truly good investment strategies because it may justify a bad strategy that pass the AIC test by coincidence. Also, we have shown that the higher the *IS* Sharpe Ratio of a model, the lower its *OOS* Sharpe Ratio. Thus, getting a high *IS* Sharpe Ratio for a single trial is not a guarantee for good returns, but more often a sign of bad *OOS* performance. To eliminate the information asymmetry and to prevent investor from getting unnecessary losses, the practitioners should make their number of trial during strategy development public. We are not telling people that backtest is a thing that we should not do, instead, we do recognize its power as *a tool of strategy validation*, but urge that it should not be *a trick in the process of strategy development.*

References

[1] Bailey, D. H., Borwein, J., Lopez de Prado, M., and Zhu, Q. J. (2014). *Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample performance.* Notices of the American Mathematical Society, 61(5), 458-471.

[2] Michie, D., Spiegelhalter, D. J., Taylor, C. C. (1994). *Machine learning.* Neural and Statistical Classification, 13(1994), 1-298.

[3] Sharpe, W. F. (1994). *The sharpe ratio.* Journal of portfolio management, 21(1), 49-58.

[4] Goodwin, T. H. (1998). *The information ratio.* Financial Analysts Journal, 54(4), 34-43.

[5] Bailey, D. H., Borwein, J., Lopez de Prado, M., Zhu, Q. J. (2016). *The probability of backtest overfitting.* Journal of Computational Finance, forthcoming.

[6] Resnick, S. I. (2013). *Extreme values, regular variation and point processes.* Springer.

[7] Embrechts, P., Klüppelberg, C., & Mikosch, T. (2013). *Modelling extremal events: for insurance and finance (Vol. 33).* Springer Science & Business Media.