

WHAT TO LOOK FOR IN A BACKTEST

Marcos López de Prado

*Lawrence Berkeley National Laboratory
Computational Research Division*



BERKELEY LAB

LAWRENCE BERKELEY NATIONAL LABORATORY



U.S. DEPARTMENT OF
ENERGY

Electronic copy available at: <https://ssrn.com/abstract=2308682>

Key Points

- Most firms and portfolio managers rely on backtests (or historical simulations of performance) to allocate capital to investment strategies.
- After trying only 7 strategy configurations, a researcher is expected to identify at least one 2-year long backtest with an annualized Sharpe ratio of over 1, when the expected out of sample Sharpe ratio is 0.
- If the researcher tries a large enough number of strategy configurations, a backtest can always be fit to any desired performance for a fixed sample length. Thus, there is a minimum backtest length (MinBTL) that should be required for a given number of trials.
- Standard statistical techniques designed to prevent regression overfitting, such as *hold-out*, are inaccurate in the context of backtest evaluation.
- The practical totality of published backtests do not report the number of trials involved.
- Under memory effects, overfitting leads to systematic losses, not noise.
- **Most backtests are overfit, and lead to losing strategies.**

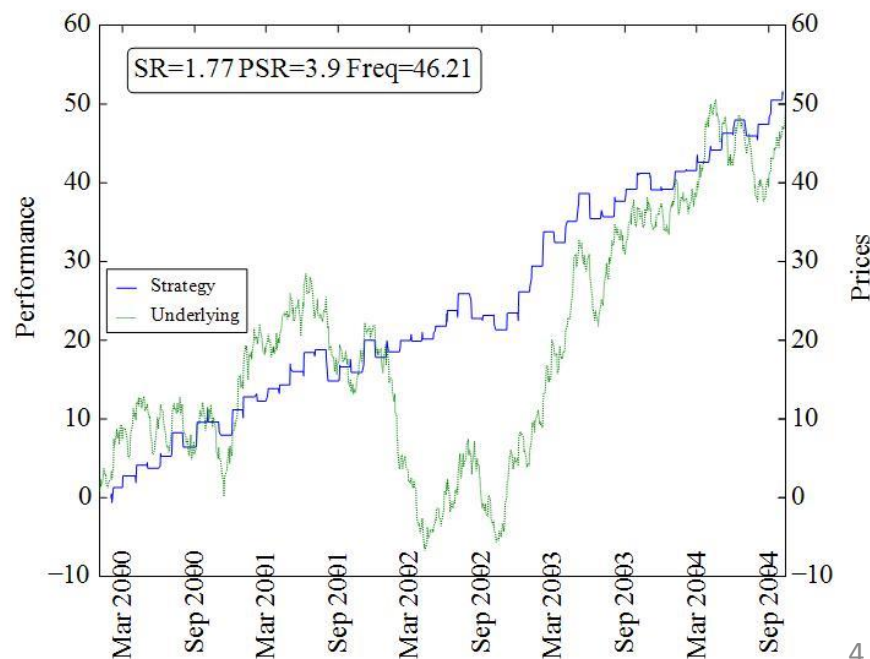
SECTION I

Backtesting Investment Strategies

Backtesting Investment Strategies

- A backtest is a historical simulation of an algorithmic investment strategy.
- Among other results, it computes the series of profits and losses that such strategy would have generated, should that algorithm had been run over that time period.

On the right, example of a backtested strategy. The green line plots the performance of a tradable security, while the blue line plots the performance achieved by buying and selling that security. Sharpe ratio is 1.77, with 46.21 trades per year. Note the low correlation between the strategy's performance and the security's.



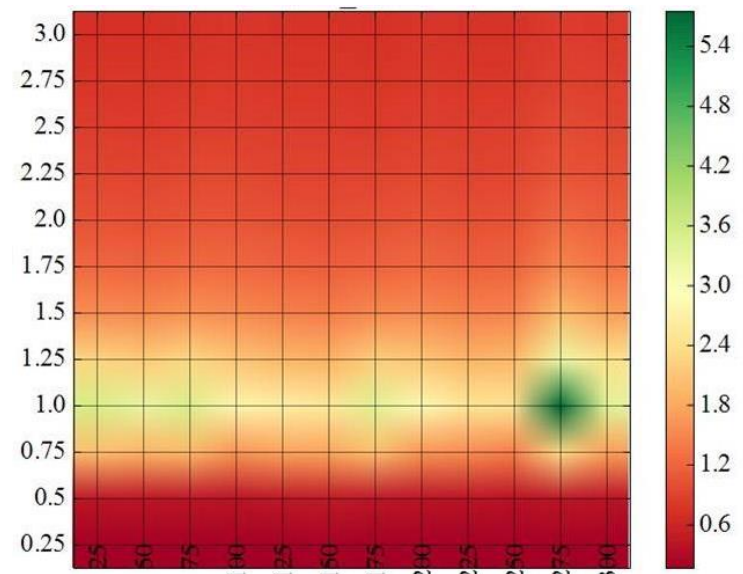
Reasons for Backtesting Investment Strategies

- The information contained in that series of profits and losses is summarized in popular performance metrics, such as the Sharpe Ratio (SR).
- These metrics are essential to decide optimal parameters combinations: Calibration frequency, risk limits, entry thresholds, stop losses, profit taking, etc.

Optimizing two parameters generates a 3D surface, which can be plotted as a heat-map.

The x-axis tries different entry thresholds, while the y-axis tries different exit thresholds.

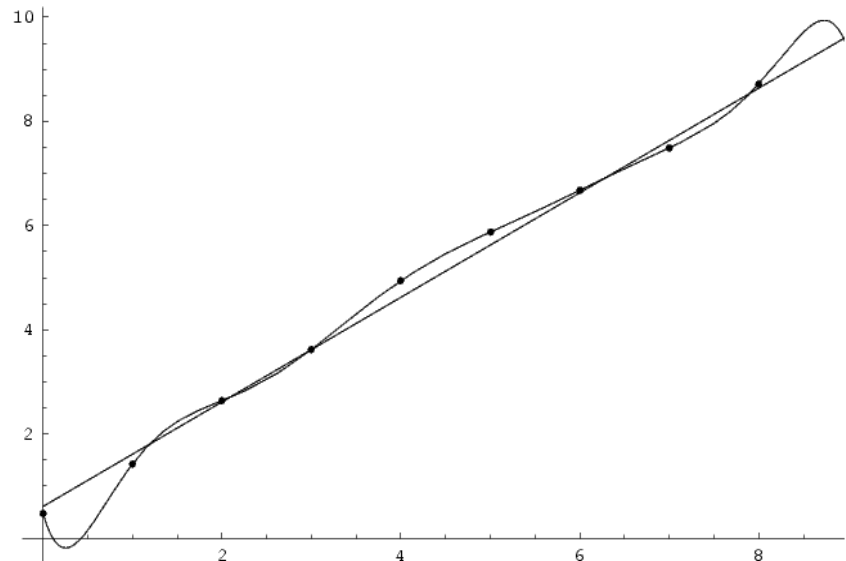
The spectrum closer to green indicates the region of optimal SR in-sample.



The Notion of Backtest Overfitting

- Given any financial series, it is relatively simple to *overfit* an investment strategy so that it performs well **in-sample (IS)**.
- Overfitting is a concept borrowed from machine learning, and denotes the situation when a model targets particular observations rather than a general structure.

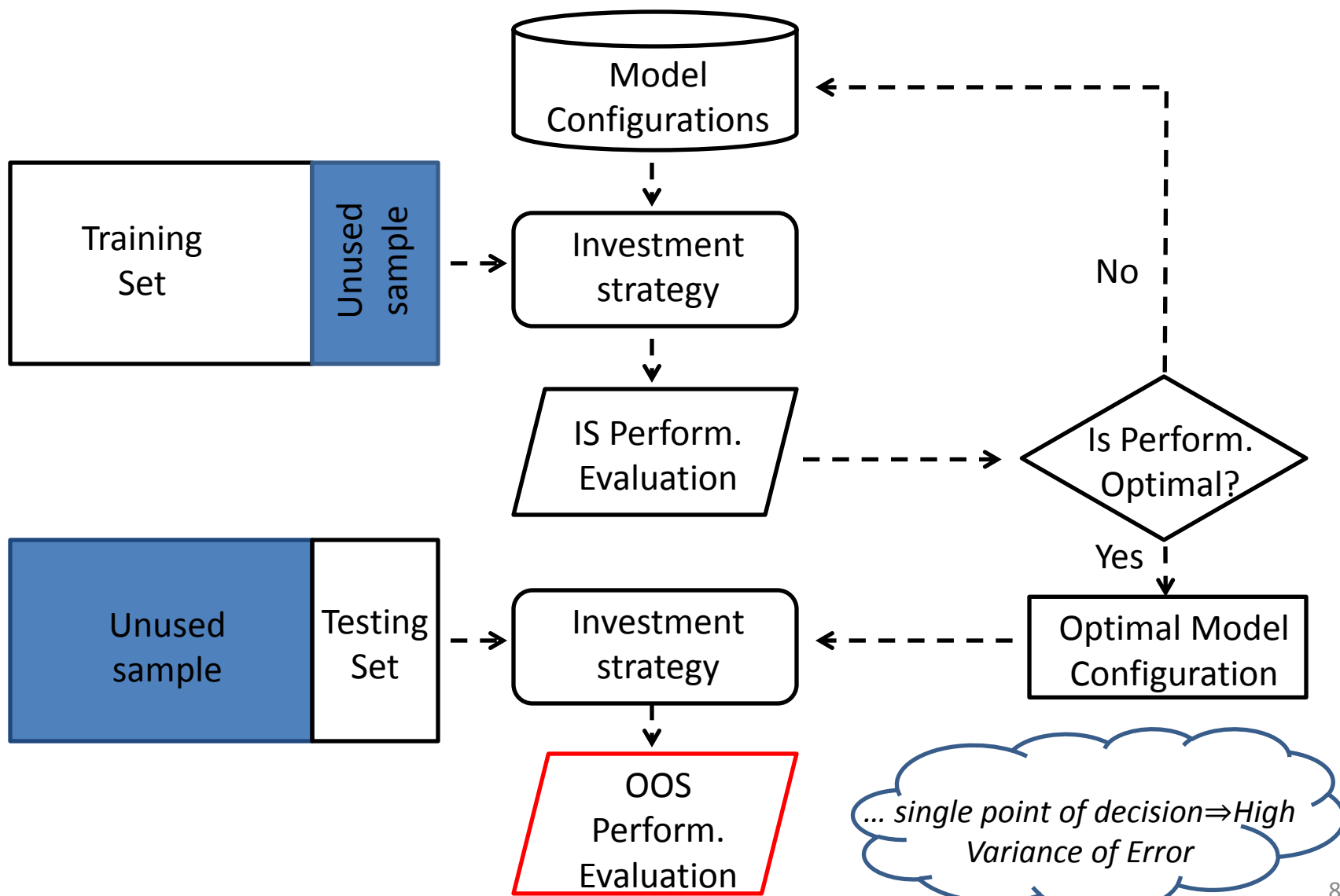
Overfitting is a well-studied problem in regression theory. This figure plots a polynomial regression vs. a linear regression. Although the former passes through every point, the simpler linear regression would produce better predictions **out-of-sample (OOS)**.



Hold-Out (1/2)

- Perhaps the most common approach among practitioners is to require the researcher to “hold-out” a part of the available sample (also called “test set” method).
- This “hold-out” is used to estimate the OOS performance, which is then compared with the IS performance.
- If they are congruent, the investor has no grounds to “reject” the hypothesis that the backtest is overfit.
- The main advantage of this procedure is its simplicity.

Hold-Out (2/2)



Why does Hold-Out fail? (1/3)

1. If the data is publicly available, the researcher may use the “hold-out” as part of the IS.
2. Even if that’s not the case, any seasoned researcher knows well how financial variables performed over the OOS interval, so that information ends up being used anyway, consciously or not.
3. Hold-out is clearly inadequate for small samples. The IS will be too short to fit, and the OOS too short to conclude anything with sufficient confidence. For example, if a strategy trades on a weekly basis, hold-out could not be used on backtests of less than 20 years (Weiss and Kulikowski [1991]).

Why does Hold-Out fail? (2/3)

4. Van Belle and Kerr [2012] point out the high variance of hold-out's estimation errors. Different “hold-outs” are likely to lead to opposite conclusions.
5. Hawkins [2004] shows that if the OOS is taken from the end of a time series, we are losing the most recent observations, which often are the most representative going forward. If the OOS is taken from the beginning of the time series, the testing will be done on the least representative portion of the data.

Why does Hold-Out fail? (3/3)

6. As long as the researcher tries more than one strategy configuration, overfitting is always present (see Section 2.1 for a proof). *The hold-out method does not take into account the number of trials attempted before selecting a model*, and consequently cannot assess the probability of backtest overfitting.

The answer to the question “is this backtest overfit?” is not a simple True or False, but a non-null Probability of Backtest Overfitting (PBO).

Later on we will show how to compute PBO.

SECTION II

How Easy is to Overfit a Backtest?

How Easy is to Overfit a Backtest? (1/3)

- **PROBLEM:** For a given strategy, a researcher would like to compare N possible model configurations, and select the configuration with optimal performance IS.
- **QUESTION #1:** How likely is she to overfit the backtest?
- *PROPOSITION #1:* Consider a set of N model configurations, each with IID Standard Normal performance. Then, a researcher is expected to find an “optimal” strategy with an IS annualized SR over y years of $E[\max_N] \approx y^{-1/2} \left((1 - \gamma)Z^{-1} \left[1 - \frac{1}{N} \right] + \gamma Z^{-1} \left[1 - \frac{1}{N} e^{-1} \right] \right)$

where γ is the Euler-Mascheroni constant, Z is the CDF of the Standard Normal and e is Euler’s number.

How Easy is to Overfit a Backtest? (2/3)

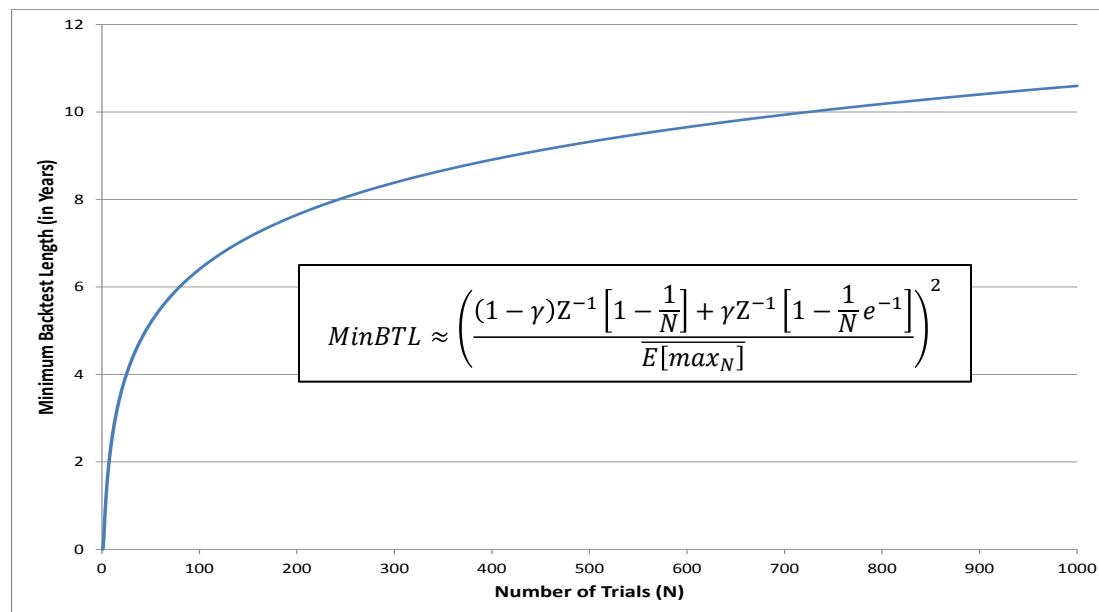
- *THEOREM #1: The Minimum Backtest Length (MinBTL, in years) needed to avoid selecting a strategy with an IS SR of $\overline{E[\max_N]}$ among N strategies with an expected OOS SR of zero is*

$$\begin{aligned} \text{MinBTL} &\approx \left(\frac{(1 - \gamma)Z^{-1} \left[1 - \frac{1}{N}\right] + \gamma Z^{-1} \left[1 - \frac{1}{N} e^{-1}\right]}{\overline{E[\max_N]}} \right)^2 \\ &< \frac{2\text{Ln}[N]}{\overline{E[\max_N]}^2} \end{aligned}$$

Note: MinBTL assesses a backtest's representativeness given N trials, while MinTRL & PSR assess a track-record's (single trial). See [Bailey and López de Prado \[2012\]](#) for further details.

How Easy is to Overfit a Backtest? (3/3)

For instance, if only 5 years of data are available, no more than 45 independent model configurations should be tried. For that number of trials, the expected maximum SR IS is 1, whereas the expected SR OOS is 0.



After trying only 7 independent strategy configurations, the expected maximum SR IS is 1 for a 2-year long backtest, while the expected SR OOS is 0.

Therefore, a backtest that does not report the number of trials N used to identify the selected configuration makes it impossible to assess the risk of overfitting.

Overfitting makes any Sharpe ratio achievable IS... the researcher just needs to keep trying alternative parameters for that strategy!!

SECTION III

The Consequences of Overfitting

Overfitting in the Absence of Memory (1/3)

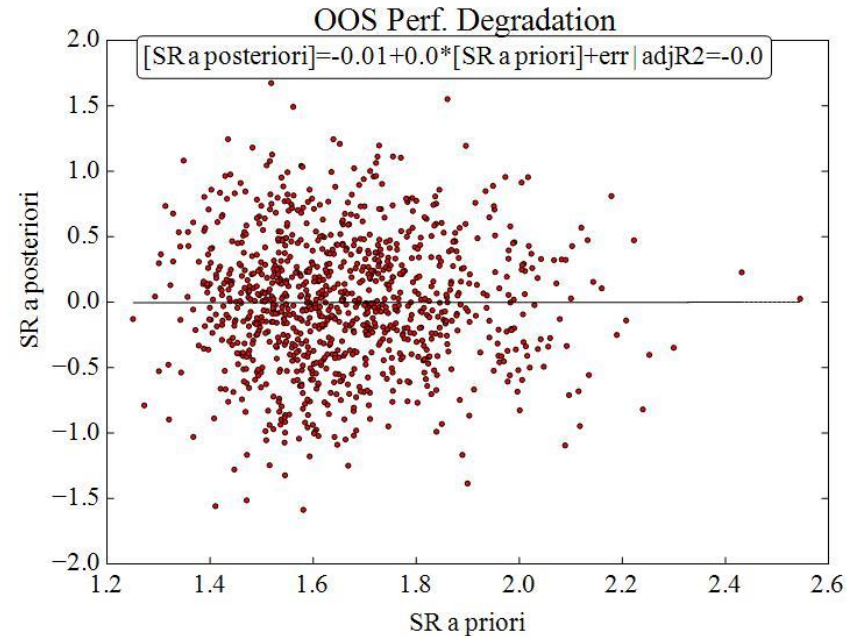
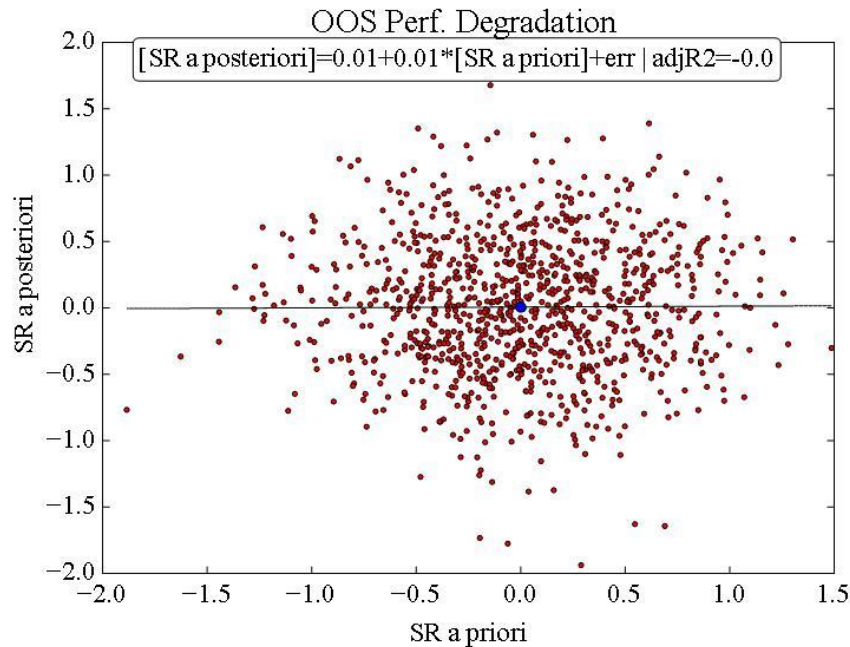
- We can generate N Gaussian random walks by drawing from a Standard Normal distribution, each walk having a size T . Each path m_τ can be obtained as a cumulative sum of Gaussian draws

$$\Delta m_\tau = \mu + \sigma \varepsilon_\tau$$

where the random shocks are IID distributed $\varepsilon_\tau \sim Z$, $\tau = 1, \dots, T$.

- We divide these paths into two disjoint samples of size $T/2$, and call the first one IS and the second one OOS.
- At the moment of choosing a particular parameter combination as optimal, the researcher had access to the IS series, not the OOS.
- **QUESTION #2: What is the relation between SR IS and SR OOS when the stochastic process has no memory?**

Overfitting in the Absence of Memory (2/3)



The left figure shows the relation between SR IS (x-axis) and SR OOS (y-axis), for $\mu = 0, \sigma = 1, N = 1000, T = 1000$. Because the process follows a random walk, the scatter plot has a circular shape centered in the point (0,0).

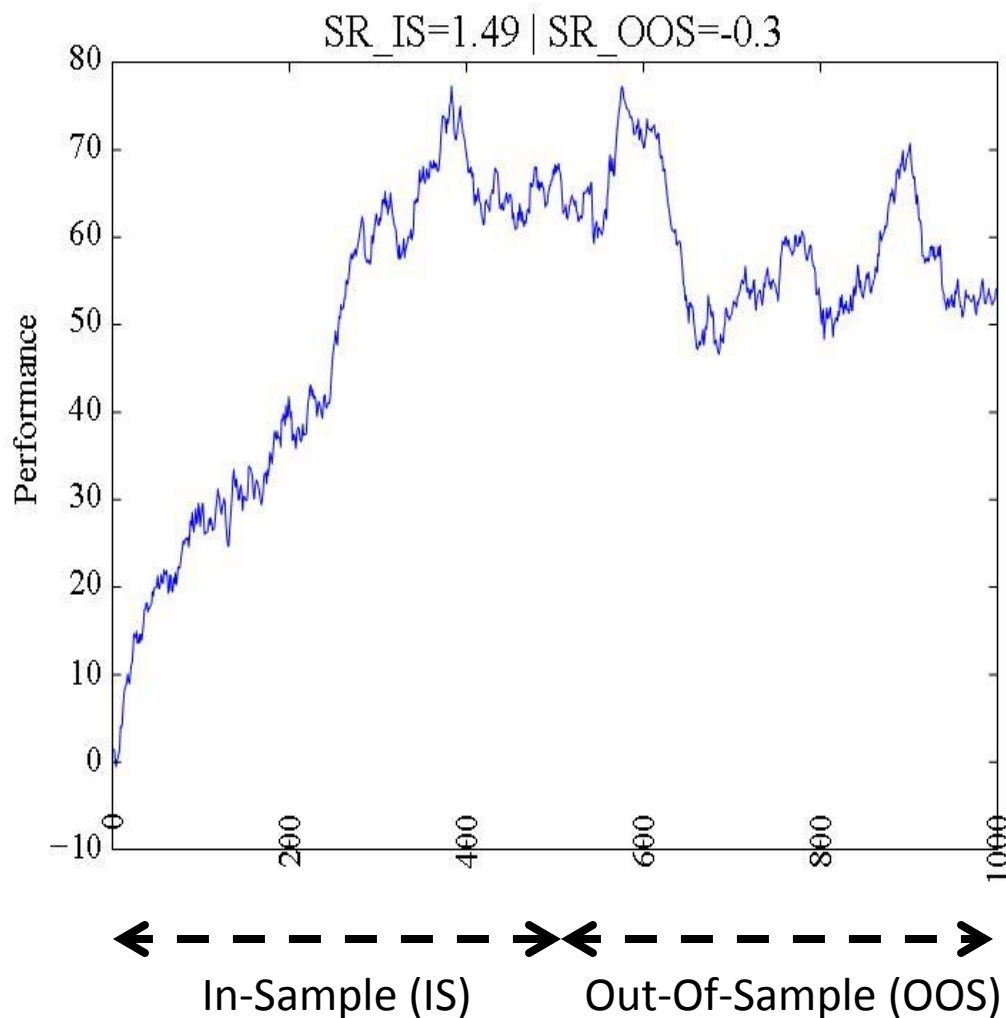
The right figure illustrates what happens once we add a “model selection” procedure. Now the SR IS ranges from 1.2 to 2.6, and it is centered around 1.7. Although the backtest for the selected model generates the expectation of a 1.7 SR, the expected SR OOS is unchanged around 0.

Overfitting in the Absence of Memory (3/3)

This figure shows what happens when we select the random walk with highest SR IS.

The performance of the first half was optimized IS, and the performance of the second half is what the investor receives OOS.

The good news is, in the absence of memory there is no reason to expect overfitting to induce negative performance.



Overfitting in the Presence of Memory (1/5)

- Unfortunately, overfitting rarely has the neutral implications discussed in the previous example, which was purposely chosen to exhibit a globally unconditional behavior.
- Centering each path to match a mean μ removes one degree of freedom.

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

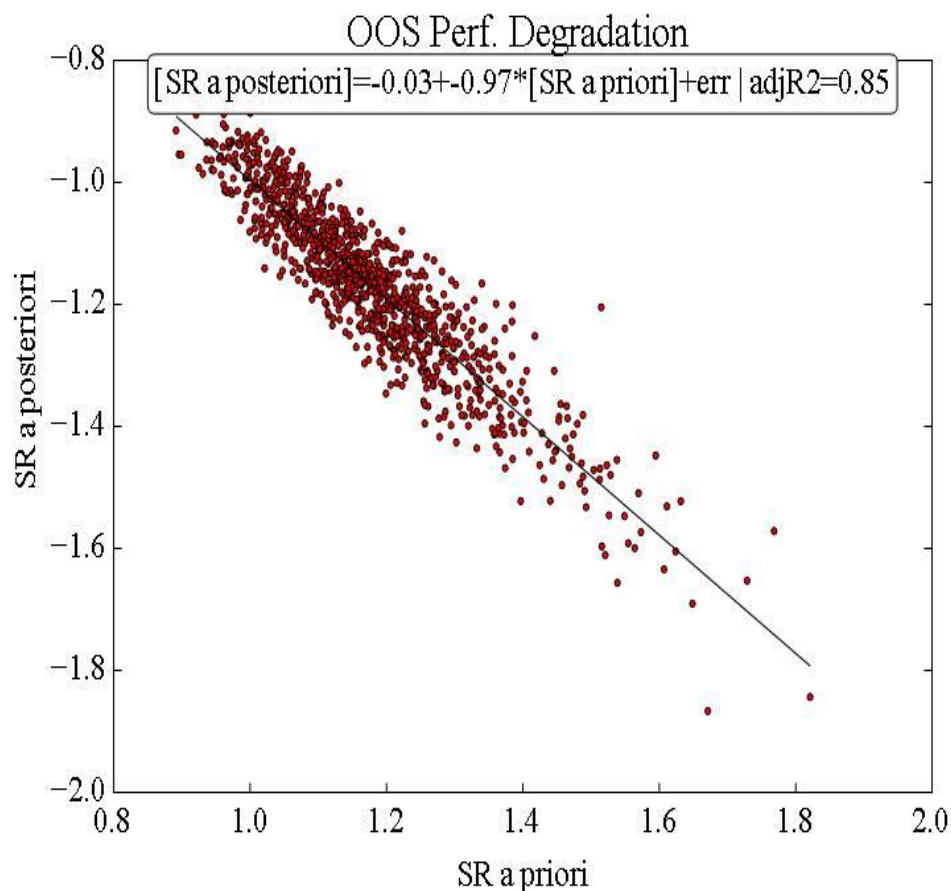
- We can re-run the same Monte Carlo experiment as before, this time on the re-centered variables $\overline{\Delta m}_\tau$.
- **QUESTION #3: What is the relation between SR IS and SR OOS when the stochastic process has memory?**

Overfitting in the Presence of Memory (2/5)

Adding this single global constraint causes the OOS performance to be negative, even though the underlying process was trendless.

Also, a strongly negative linear relation between performance IS and OOS arises, **indicating that the more we optimize IS, the worse is OOS performance.**

The p-values associated with the intercept and the IS performance (SR a priori) are respectively 0.5005 and 0, indicating that the negative linear relation between IS and OOS Sharpe ratios is statistically significant.



Overfitting in the Presence of Memory (3/5)

- PROPOSITION 2: Given two alternative configurations (A and B) of the same model, where $\sigma_{IS}^A = \sigma_{OOS}^A = \sigma_{IS}^B = \sigma_{OOS}^B$, imposing a global constraint $\mu^A = \mu^B$ implies that

$$SR_{IS}^A > SR_{IS}^B \Leftrightarrow SR_{OOS}^A < SR_{OOS}^B$$

- Another way of introducing memory is through serial-conditionality, like in a first-order autoregressive process.

$$\begin{aligned}\Delta m_\tau &= (1 - \varphi)\mu + (\varphi - 1)m_{\tau-1} + \sigma\varepsilon_\tau \\ m_\tau &= (1 - \varphi)\mu + \varphi m_{\tau-1} + \sigma\varepsilon_\tau\end{aligned}$$

where the random shocks are IID distributed as $\varepsilon_\tau \sim Z$.

Overfitting in the Presence of Memory (4/5)

- PROPOSITION 3: The half-life period of a first-order autoregressive process with autoregressive coefficient $\varphi \in (0,1)$ occurs at

$$\tau = -\frac{\text{Ln}[2]}{\text{Ln}[\varphi]}$$

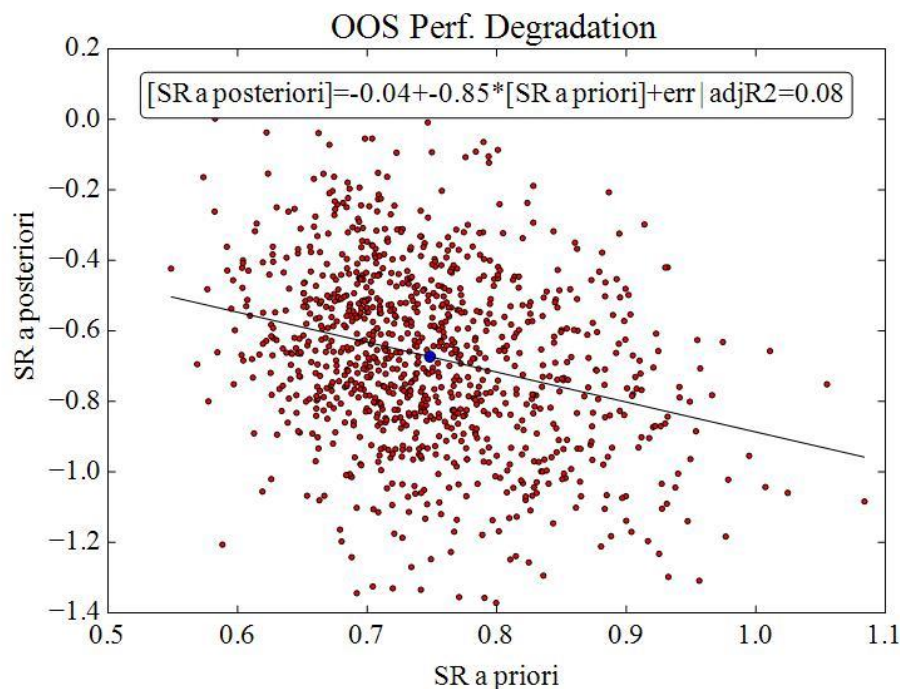
- PROPOSITION 4: Given two alternative configurations (A and B) of the same model, where $\sigma_{IS}^A = \sigma_{OOS}^A = \sigma_{IS}^B = \sigma_{OOS}^B$ and the P&L series follows the same first-order autoregressive stationary process,

$$SR_{IS}^A > SR_{IS}^B \Leftrightarrow SR_{OOS}^A < SR_{OOS}^B$$

- Proposition 4 reaches the same conclusion as Proposition 2 (a compensation effect), without requiring a global constraint.

Overfitting in the Presence of Memory (5/5)

For example, if $\varphi = 0.995$, it takes about 138 observations to retrace half of the deviation from the equilibrium. This introduces another form of compensation effect, just as we saw in the case on a global constraint. We have re-run the previous Monte Carlo experiment, this time on the autoregressive process with $\mu = 0, \sigma = 1, \varphi = 0.995$, and plotted the pairs of performance IS vs. OOS.



Because financial time series are known to exhibit memory (in the form of economic cycles, reversal of financial flows, structural breaks, bubbles' bursts, etc.), the consequence of overfitting is negative performance out-of-sample.

SECTION IV

Combinatorially-Symmetric Cross-Validation

A formal definition of Backtest Overfitting

- **QUESTION #4: What is the probability that an “optimal” strategy is overfit?**
- DEFINITION 1 (Overfitting): Let be n^* the strategy with optimal performance IS, i.e. $R_{n^*} \geq R_n, \forall n = 1, \dots, N$. Denote \overline{R}_{n^*} the performance OOS of n^* . Let be $Me[\overline{R}]$ the median performance of all strategies OOS. Then, we say that a strategy selection process overfits if for a strategy n^* with the highest rank IS,

$$E[\overline{R}_{n^*}] < Me[\overline{R}]$$

- In the above definition we refer to overfitting in relation to the strategy selection process (e.g., backtesting), not a strategy’s model calibration (e.g., a regression).

A formal definition of PBO

- DEFINITION 2 (Probability of Backtest Overfitting): Let be n^* the strategy with optimal performance IS. Because strategy n^* is not necessarily optimal OOS, there is a non-null probability that $\overline{R_{n^*}} < Me[\overline{R}]$. We define the probability that the selected strategy n^* is overfit as

$$PBO \equiv Prob \left[\overline{R_{n^*}} < Me[\overline{R}] \right]$$

- In other words, we say that a strategy selection process overfits if the expected performance of the strategies selected IS is less than the median performance OOS of all strategies. **In that situation, the strategy selection process becomes in fact detrimental.**

Combinatorially-Symmetric Cross-Validation (1/4)

1. Form a matrix ***M*** by collecting the performance series from the *N* trials.
2. Partition ***M*** across rows, into an even number *S* of disjoint submatrices of equal dimensions. Each of these submatrices ***M_s***, with *s=1,...,S*, is of order $\left(\frac{T}{S} \times N\right)$.
3. Form all combinations *C_S* of ***M_s***, taken in groups of size $\frac{S}{2}$.
This gives a total number of combinations

$$\binom{S}{S/2} = \binom{S-1}{S/2-1} \frac{S}{S/2} = \cdots = \prod_{i=0}^{S/2-1} \frac{S-i}{S/2-i}$$

Combinatorially-Symmetric Cross-Validation (2/4)

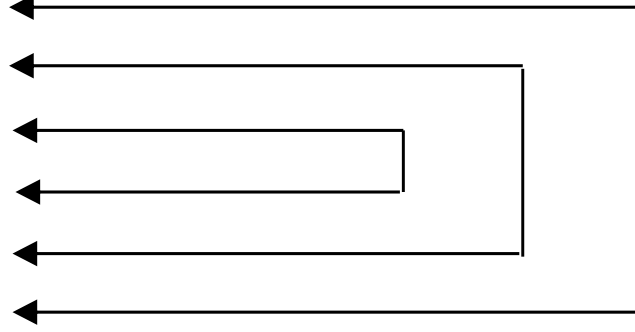
4. For each combination $c \in C_S$,
 - a. Form the *training set* \mathbf{J} , by joining the $S/2$ submatrices \mathbf{M}_s that constitute c . \mathbf{J} is a matrix of order $\left(\frac{T}{S} \frac{S}{2} \times N\right) = \left(\frac{T}{2} \times N\right)$.
 - b. Form the *testing set* $\bar{\mathbf{J}}$, as the complement of \mathbf{J} in \mathbf{M} . In other words, $\bar{\mathbf{J}}$ is the $\left(\frac{T}{2} \times N\right)$ matrix formed by all rows of \mathbf{M} that are not part of \mathbf{J} .
 - c. Form a vector \mathbf{R} of performance statistics of order N , where the n -th item of \mathbf{R} reports the performance associated with the n -th column of \mathbf{J} (the training set).
 - d. Determine the element n^* such that $R_n \leq R_{n^*}, \forall n = 1, \dots, N$. In other words, $n^* = \arg \max_n \{R_n\}$.

Combinatorially-Symmetric Cross-Validation (3/4)

4. (... continuation.)
- e. Form a vector $\overline{\mathbf{R}}$ of performance statistics of order N , where the n -th item of $\overline{\mathbf{R}}$ reports the performance associated with the n -th column of $\overline{\mathbf{J}}$ (the testing set).
- f. Determine the relative rank of $\overline{R_{n^*}}$ within $\overline{\mathbf{R}}$. We will denote this relative rank as $\overline{\omega}_c$, where $\overline{\omega}_c \in (0,1)$. This is the relative rank of the OOS performance associated with the trial chosen IS. If the strategy optimization procedure is not overfitting, we should observe that $\overline{R_{n^*}}$ systematically outperforms $\overline{\mathbf{R}}$ OOS, just as R_{n^*} outperformed \mathbf{R} .
- g. We define the logit $\lambda_c = \text{Ln} \frac{\overline{\omega}_c}{1-\overline{\omega}_c}$. This presents the property that $\lambda_c = 0$ when $\overline{R_{n^*}}$ coincides with the median of $\overline{\mathbf{R}}$.

Combinatorially-Symmetric Cross-Validation (4/4)

IS		OOS	
A	B	C	D
A	C	B	D
A	D	B	C
B	C	A	D
B	D	A	C
C	D	A	B



The diagram shows a table with two main columns: 'IS' (In-Sample) and 'OOS' (Out-of-Sample). Each row represents a unique combination of four elements (A, B, C, D). The 'IS' column contains two elements, and the 'OOS' column contains two elements. Arrows point from the 'OOS' column to the 'IS' column for each row, indicating the mapping of training and testing sets. The arrows are as follows: Row 1 (A, B | C, D) has an arrow from C to A; Row 2 (A, C | B, D) has an arrow from B to A; Row 3 (A, D | B, C) has an arrow from B to A; Row 4 (B, C | A, D) has an arrow from A to B; Row 5 (B, D | A, C) has an arrow from A to B; Row 6 (C, D | A, B) has an arrow from A to C.

This figure schematically represents how the combinations in C_S are used to produce training and testing sets, where $S=4$. Each arrow is associated with a logit, λ_c .

5. Compute the distribution of ranks OOS by collecting all the logits λ_c , for $c \in C_S$. $f(\lambda)$ is then the relative frequency at which λ occurred across all C_S , with $\int_{-\infty}^{\infty} f(\lambda) d\lambda = 1$.

SECTION V

Assessing the Representativeness of a Backtest

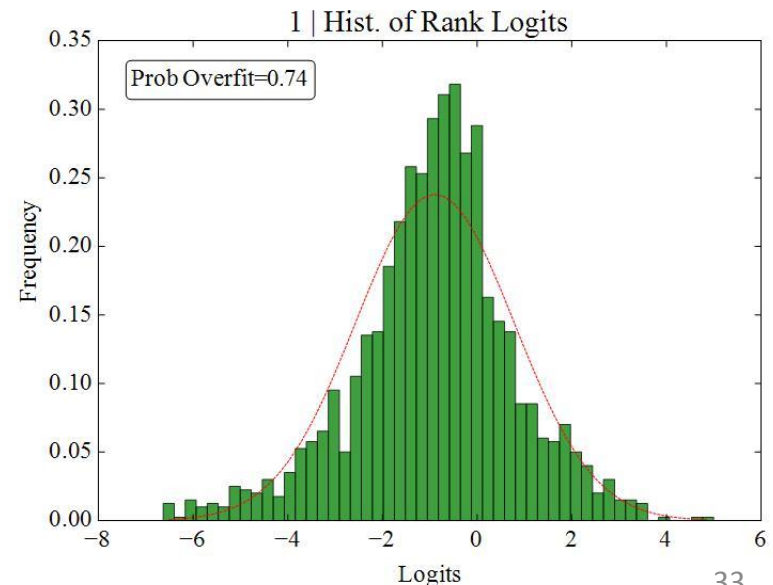
Tool #1: Prob. of Backtest Overfitting (PBO)

- PBO was defined earlier as $Prob \left[\overline{R}_{n^*} < Me[\overline{R}] \right]$.
- The framework described in the previous section has given us the tools to estimate PBO as

$$\phi = \int_{-\infty}^0 f[\lambda] d\lambda$$

This represents the rate at which optimal IS strategies underperform the median of the OOS trials.

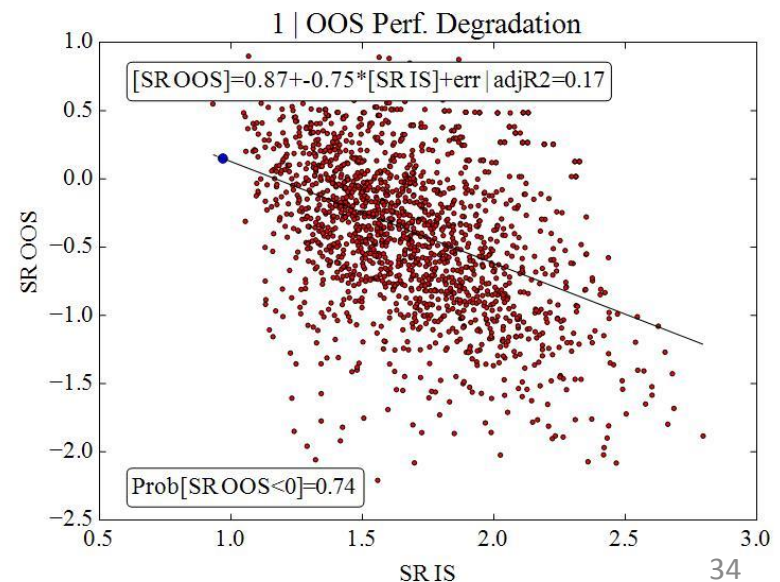
The analogue of \overline{R} in medical research is the placebo given to a portion of patients in the test set. If the backtest is truly helpful, the optimal strategy selected IS should outperform most of the N trials OOS ($\lambda_c > 0$).



Tool #2: Perform. Degradation and Prob. of Loss

- The previous section introduced the procedure to compute, among other results, the pair $(R_n^*, \overline{R_n^*})$ for each combination $c \in \mathcal{C}_S$.
- The pairs $(R_n^*, \overline{R_n^*})$ allow us to visualize how strong is the performance degradation, and obtain a more realistic range of attainable performance OOS.

A particularly useful statistic is the proportion of combinations with negative performance, $\text{Prob} \left[\overline{R_n^*} < 0 \right]$. Note that, even if $\phi < \frac{1}{2}$, $\text{Prob} \left[\overline{R_n^*} < 0 \right]$ could be high, in which case the strategy's performance OOS is poor for reasons other than overfitting.

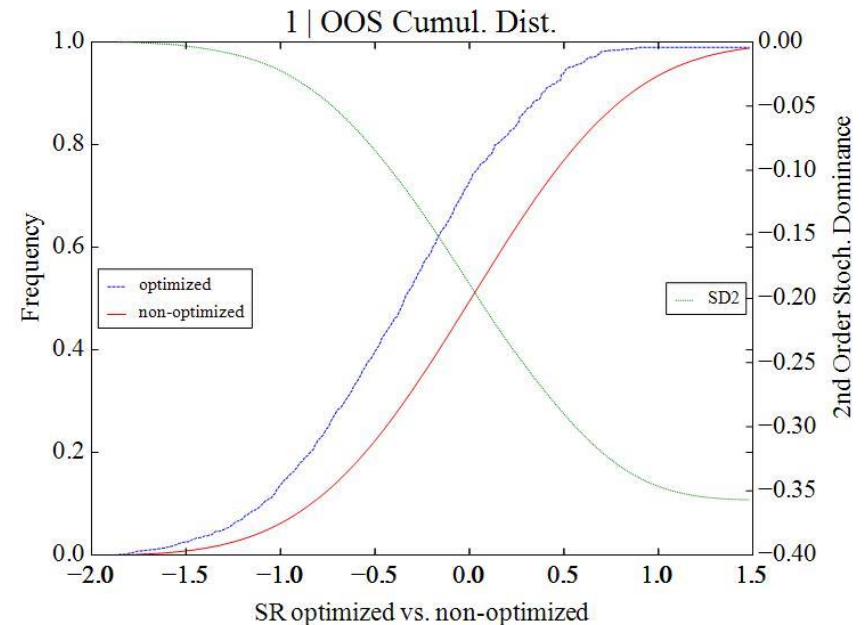


Tool #3: Stochastic Dominance

- Stochastic dominance allows us to rank gambles or lotteries without having to make strong assumptions regarding an individual's utility function.

In the context of our framework, first-order stochastic dominance occurs if $Prob[\overline{R}_{n^*} \geq x] \geq Prob[\overline{R} \geq x]$ for all x , and for some x , $Prob[\overline{R}_{n^*} \geq x] > Prob[\overline{R} \geq x]$.

A less demanding criterion is second-order stochastic dominance: $SD2[x] = \int_{-\infty}^x (Prob[\overline{R} \leq x] - Prob[\overline{R}_{n^*} \leq x])dx \geq 0$ for all x , and that $SD2[x] > 0$ at some x .



SECTION VI

Features and Accuracy of CSCV's Estimates

Features of CSCV (1/2)

1. CSCV ensures that the training and testing sets are of equal size, thus providing comparable accuracy to the IS and OOS Sharpe ratios (or any performance metric susceptible to sample size).
2. CSCV is *symmetric*, in the sense that all training sets are re-used as testing sets. In this way, the decline in performance can only result from overfitting, not discrepancies between the training and testing sets.
3. CSCV respects the time-dependence and other seasonalities present in the data, because it does not require a random allocation of the observations to the S subsamples.
4. CSCV derives a non-random distribution of logits, in the sense that each logit is deterministically derived from one item in the set of combinations C_S . Multiple runs of CSCV return the same ϕ , which can be independently replicated and verified by another user.

Features of CSCV (2/2)

5. The dispersion of the distribution of logits conveys relevant information regarding the robustness of the strategy selection procedure. A robust strategy selection leads to a consistent OOS performance rankings, which translate into similar logits.
6. Our procedure to estimate PBO is model-free, in the sense that it does not require the researcher to specify a forecasting model or the definitions of forecasting errors.
7. It is also non-parametric, as we are not making distributional assumptions on PBO. This is accomplished by using the concept of logit, λ_c . If good backtesting results are conducive to good OOS performance, the distribution of logits will be centered in a significantly positive value, and its left tail will marginally cover the region of negative logit values, making $\phi \approx 0$.

CSCV Accuracy via Monte Carlo

First, we have computed CSCV's PBO on 1,000 randomly generated matrices \mathbf{M} for every parameter combination (\widetilde{SR}, T, N) . This has provided us with 1,000 independent estimates of PBO for every parameter combination, with a mean and standard deviation reported in columns Mean_CSCV and Std_CSCV.

Second, we generated 1,000 matrices \mathbf{M} (experiments) for various test cases of order $(T \times N) = (1000 \times 100)$, and computed the proportion of experiments that yielded an OOS performance below the median. The proportion of IS optimal selections that underperformed OOS is reported in Prob_MC. This Prob_MC is well within the confidence bands implied by Mean_CSCV and Std_CSCV.

SR_Case	T	N	Mean_CSCV	Std_CSCV	Prob_MC	CSCV-MC
0	500	500	1.000	0.000	1.000	0.000
0	1000	500	1.000	0.000	1.000	0.000
0	2500	500	1.000	0.000	1.000	0.000
0	500	100	1.000	0.000	1.000	0.000
0	1000	100	1.000	0.000	1.000	0.000
0	2500	100	1.000	0.000	1.000	0.000
0	500	50	1.000	0.000	1.000	0.000
0	1000	50	1.000	0.000	1.000	0.000
0	2500	50	1.000	0.000	1.000	0.000
0	500	10	1.000	0.001	1.000	0.000
0	1000	10	1.000	0.000	1.000	0.000
0	2500	10	1.000	0.000	1.000	0.000
1	500	500	0.993	0.007	0.991	0.002
1	1000	500	0.893	0.032	0.872	0.021
1	2500	500	0.561	0.022	0.487	0.074
1	500	100	0.929	0.023	0.924	0.005
1	1000	100	0.755	0.034	0.743	0.012
1	2500	100	0.371	0.034	0.296	0.075
1	500	50	0.870	0.031	0.878	-0.008
1	1000	50	0.666	0.035	0.628	0.038
1	2500	50	0.288	0.047	0.199	0.089
1	500	10	0.618	0.054	0.650	-0.032
1	1000	10	0.399	0.054	0.354	0.045
1	2500	10	0.123	0.048	0.093	0.030
2	500	500	0.679	0.037	0.614	0.065
2	1000	500	0.301	0.038	0.213	0.088
2	2500	500	0.011	0.011	0.000	0.011
2	500	100	0.488	0.035	0.413	0.075
2	1000	100	0.163	0.045	0.098	0.065
2	2500	100	0.004	0.006	0.002	0.002
2	500	50	0.393	0.040	0.300	0.093
2	1000	50	0.113	0.044	0.068	0.045
2	2500	50	0.002	0.004	0.000	0.002
2	500	10	0.186	0.054	0.146	0.040
2	1000	10	0.041	0.027	0.011	0.030
2	2500	10	0.000	0.001	0.000	0.000
3	500	500	0.247	0.043	0.174	0.073
3	1000	500	0.020	0.017	0.005	0.015
3	2500	500	0.000	0.000	0.000	0.000
3	500	100	0.124	0.042	0.075	0.049
3	1000	100	0.007	0.008	0.001	0.006
3	2500	100	0.000	0.000	0.000	0.000
3	500	50	0.088	0.037	0.048	0.040
3	1000	50	0.004	0.006	0.002	0.002
3	2500	50	0.000	0.000	0.000	0.000
3	500	10	0.028	0.022	0.010	0.018
3	1000	10	0.001	0.002	0.000	0.001
3	2500	10	0.000	0.000	0.000	0.000

CSCV Accuracy via Extreme Value Theory (1/3)

- The Gaussian distribution belongs to the Maximum Domain of Attraction of the Gumbel distribution, thus $\max_N \sim \Lambda[\alpha, \beta]$, where α, β are the normalizing constants and Λ is the CDF of the Gumbel distribution.
- It is known that the mean and standard deviation of a Gumbel distribution are $E[\max_N] = \alpha + \gamma\beta$, $\sigma[\max_N] = \frac{\beta\pi}{\sqrt{6}}$, where γ is the Euler-Mascheroni constant.
- Applying the method of moments, we can derive:
 - Given an estimate of $\hat{\sigma}[\max_N]$, $\hat{\beta} = \frac{\hat{\sigma}[\max_N]\sqrt{6}}{\pi}$.
 - Given an estimate of $\hat{E}[\max_N]$, and the previously obtained $\hat{\beta}$, we can estimate $\hat{\alpha} = \hat{E}[\max_N] - \gamma\hat{\beta}$.

CSCV Accuracy via Extreme Value Theory (2/3)

- These parameters allow us to model the distribution of the maximum Sharpe ratio IS out of a set of $N-1$ trials. PBO can then be directly computed as $\phi = \phi_1 + \phi_2$, where:

$$\phi_1 = \int_{-\infty}^{2\widetilde{SR}} N \left[SR, \widetilde{SR}, \frac{1 + \frac{1}{2}\widetilde{SR}^2}{T} \right] (1 - \Lambda[\max(0, SR), \alpha, \beta]) dSR$$

$$\phi_2 = \int_{2\widetilde{SR}}^{\infty} N \left[SR, \widetilde{SR}, \frac{1 + \frac{1}{2}\widetilde{SR}^2}{T} \right] dSR$$

Probability ϕ_1 accounts for selecting IS a strategy with $SR_n = 0$, as a result of $SR_{N,IS} < SR_{IS}^*$. The integral has an upper boundary in $2\widetilde{SR}$ because beyond that point all trials lead to $SR_{OOS} < Me[SR_{OOS}]$, including the N th trial. That probability is accounted for by ϕ_2 , which has a lower boundary of integration in $2\widetilde{SR}$.

CSCV Accuracy via Extreme Value Theory (2/3)

A comparison of the Mean_CSCV probability with the EVT result gives us an average absolute error is 2.1%, with a standard deviation of 2.9%. The maximum absolute error is 9.9%. That occurred for the combination $(\widetilde{SR}, T, N) = (3, 500, 500)$, whereby CSCV gave a more conservative estimate (24.7% instead of 14.8%). There is only one case where CSCV underestimated PBO, with an absolute error of 0.1%. The median error is only 0.7%, with a 5%-tile of 0% and a 95%-tile of 8.51%.

In conclusion, CSCV provides accurate estimates of PBO, with relatively small errors on the conservative side.

SR_Case	T	N	Mean_CSCV	Std_CSCV	Prob_EVT	CSCV-EVT
0	500	500	1.000	0.000	1.000	0.000
0	1000	500	1.000	0.000	1.000	0.000
0	2500	500	1.000	0.000	1.000	0.000
0	500	100	1.000	0.000	1.000	0.000
0	1000	100	1.000	0.000	1.000	0.000
0	2500	100	1.000	0.000	1.000	0.000
0	500	50	1.000	0.000	1.000	0.000
0	1000	50	1.000	0.000	1.000	0.000
0	2500	50	1.000	0.000	1.000	0.000
0	500	10	1.000	0.001	1.000	0.000
0	1000	10	1.000	0.000	1.000	0.000
0	2500	10	1.000	0.000	1.000	0.000
1	500	500	0.993	0.007	0.994	-0.001
1	1000	500	0.893	0.032	0.870	0.023
1	2500	500	0.561	0.022	0.476	0.086
1	500	100	0.929	0.023	0.926	0.003
1	1000	100	0.755	0.034	0.713	0.042
1	2500	100	0.371	0.034	0.288	0.083
1	500	50	0.870	0.031	0.859	0.011
1	1000	50	0.666	0.035	0.626	0.041
1	2500	50	0.288	0.047	0.220	0.068
1	500	10	0.618	0.054	0.608	0.009
1	1000	10	0.399	0.054	0.360	0.039
1	2500	10	0.123	0.048	0.086	0.036
2	500	500	0.679	0.037	0.601	0.079
2	1000	500	0.301	0.038	0.204	0.097
2	2500	500	0.011	0.011	0.002	0.009
2	500	100	0.488	0.035	0.405	0.084
2	1000	100	0.163	0.045	0.099	0.065
2	2500	100	0.004	0.006	0.001	0.003
2	500	50	0.393	0.040	0.312	0.081
2	1000	50	0.113	0.044	0.066	0.047
2	2500	50	0.002	0.004	0.000	0.002
2	500	10	0.186	0.054	0.137	0.049
2	1000	10	0.041	0.027	0.023	0.018
2	2500	10	0.000	0.001	0.000	0.000
3	500	500	0.247	0.043	0.148	0.099
3	1000	500	0.020	0.017	0.005	0.015
3	2500	500	0.000	0.000	0.000	0.000
3	500	100	0.124	0.042	0.068	0.056
3	1000	100	0.007	0.008	0.002	0.005
3	2500	100	0.000	0.000	0.000	0.000
3	500	50	0.088	0.037	0.045	0.043
3	1000	50	0.004	0.006	0.001	0.003
3	2500	50	0.000	0.000	0.000	0.000
3	500	10	0.028	0.022	0.015	0.013
3	1000	10	0.001	0.002	0.001	0.000
3	2500	10	0.000	0.000	0.000	0.000

SECTION VI

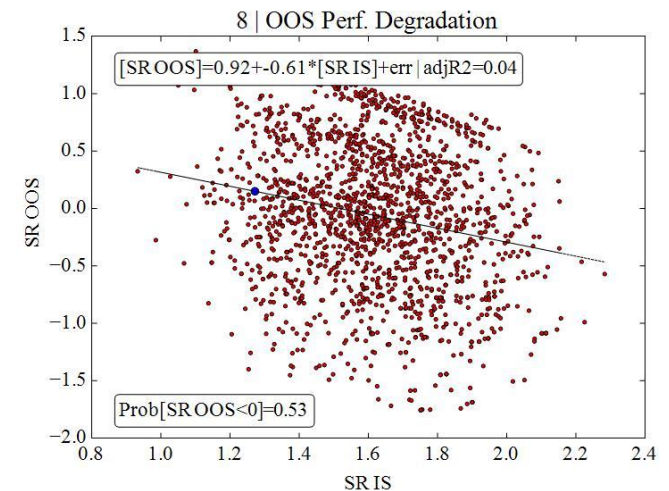
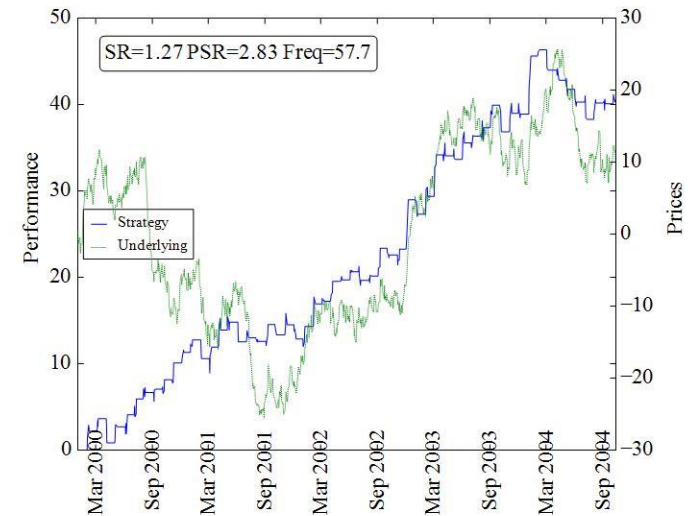
A Practical Application

A practical Application: Seasonal Effects

- There is a large number of instances where asset managers engage in predictable actions on a calendar basis. It comes as no surprise that a very popular investment strategy among hedge funds is to profit from such seasonal effects.
- Suppose that we would like to identify the optimal monthly trading rule, given four parameters:
 - **Entry_day**: Determines the business day of the month when we enter a position.
 - **Holding_period**: Gives the number of days that the position is held.
 - **Stop_loss**: Determines the size of the loss, as a multiple of the series' volatility, which triggers an exit for that month's position.
 - **Side**: Defines whether we will hold long or short positions on a monthly basis.

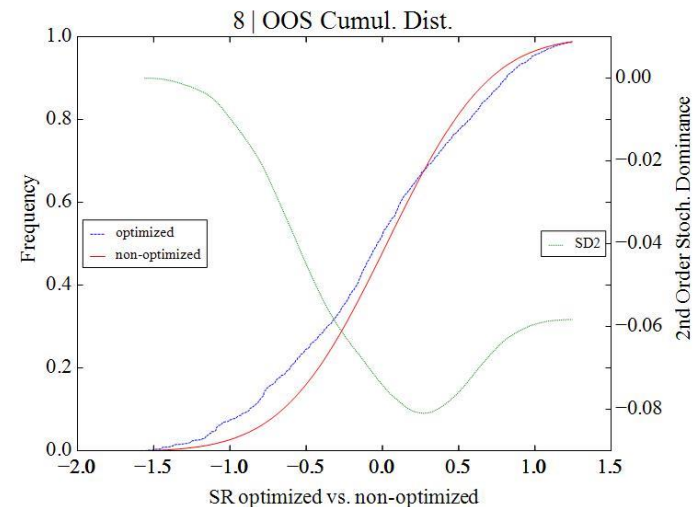
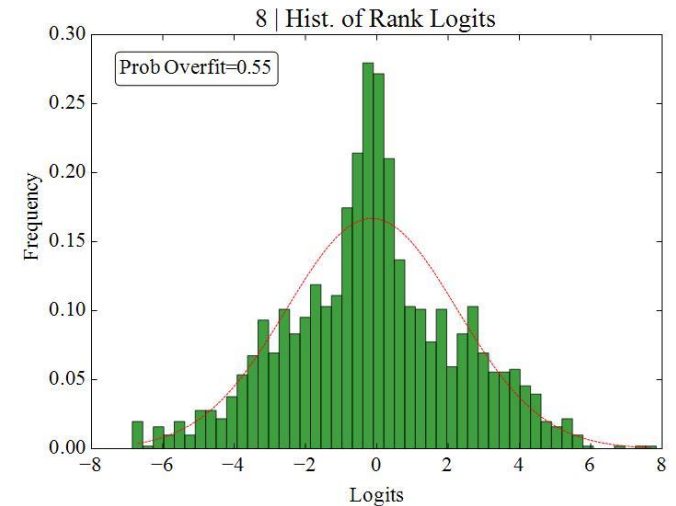
Backtest in Absence of a Seasonal Effect (1/2)

- We have generated a time series of 1000 daily prices (about 4 years), following a random walk.
- The PSR-Stat of the optimal model configuration is 2.83, which implies a less than 1% probability that the true Sharpe ratio is below 0.
- SR OOS of optimal configurations is negative in 53% of cases.
- **We have been able to identify a seasonal strategy with a SR of 1.27 despite the fact that no seasonal effect exists!!**



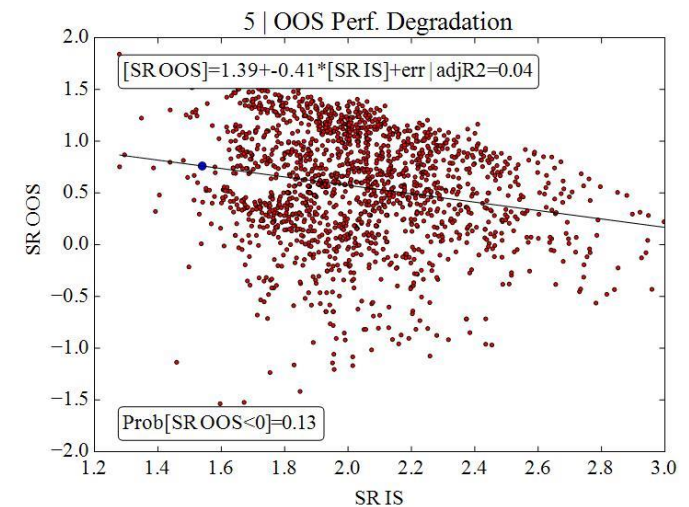
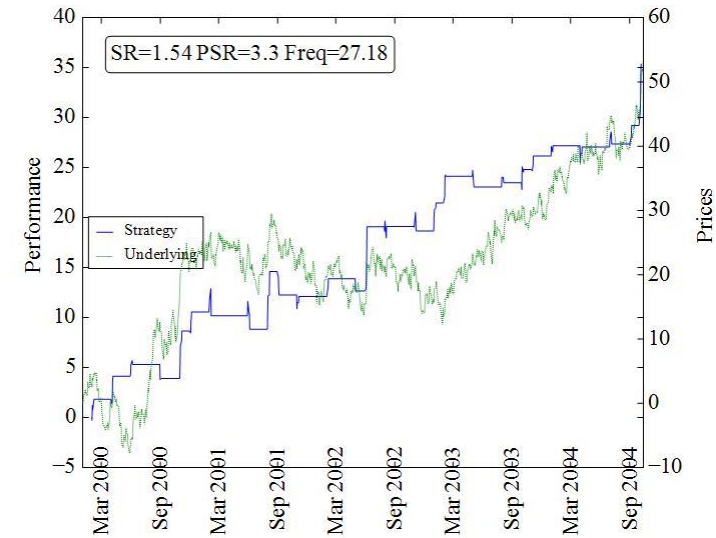
Backtest in Absence of a Seasonal Effect (2/2)

- The distribution of logits implies that, despite the elevated SR IS, the PBO is as high as 55%.
- Consequently, the distribution of optimized OOS SR does not dominate the overall distribution of OOS SR.
- **The CSCV analysis has succeeded in rejecting the overfit backtest.**



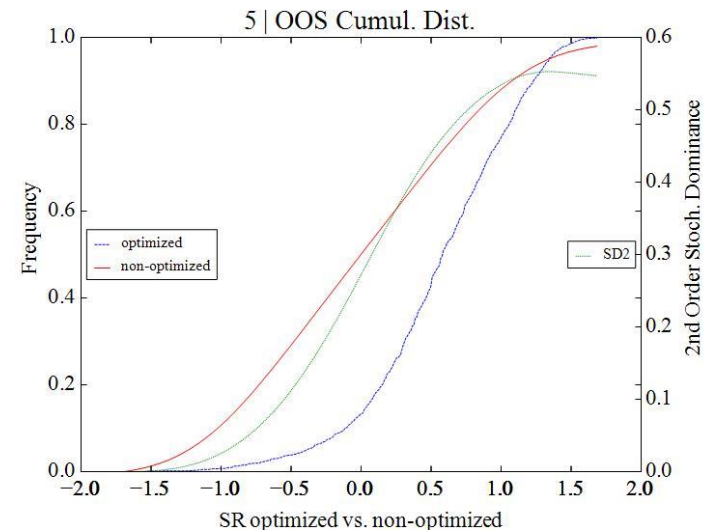
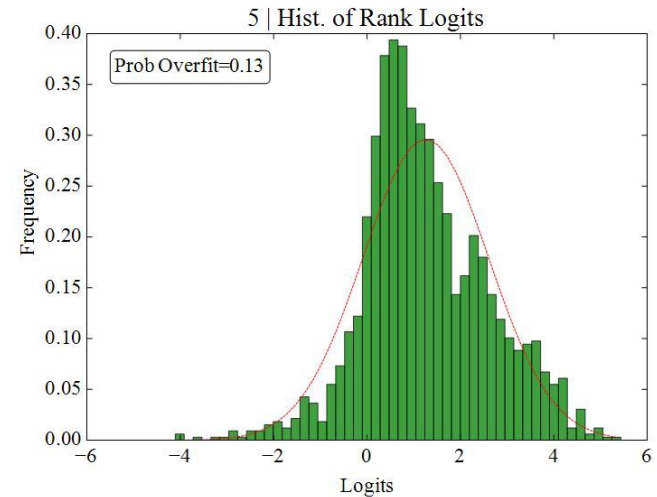
Backtest in Presence of a Seasonal Effect (1/2)

- We have taken the previous 1000 series and shifted the returns of the first 5 observations of each month by a quarter of a standard deviation.
- This generates a monthly seasonal effect, which our strategy selection procedure should discover.
- The Sharpe Ratio is similar to the previous (overfit) case (1.5 vs. 1.3).
- However, the SR OOS of optimal configurations is negative in only 13% of cases (compared to 53%).



Backtest in Presence of a Seasonal Effect (2/2)

- The distribution of logits implies that the PBO is only 13%.
- Consistently, the distribution of optimized OOS SR dominates (in first and second order) the overall distribution of OOS SR.
- **The CSCV analysis has correctly recognized the validity of this backtest, in the sense that performance inflation from overfitting is small.**



SECTION VII

Conclusions

Conclusions (1/2)

1. Backtest overfitting is difficult to avoid.
2. For a sufficiently large number of trials, it is trivial to achieve any desired Sharpe ratio for a backtest.
3. Given that most published backtests do not report the number of trials attempted, we must suppose that many of them are overfit.
4. In that case, if an investor allocates capital to those strategies, OOS performance will vary:
 - If the process has no memory: Performance will be around zero.
 - If the process has memory: Performance will be (very) negative.
5. **We suspect that backtest overfitting is a leading reason why so many algorithmic or systematic hedge funds fail.**

Conclusions (2/2)

6. Standard statistical techniques designed to detect overfitting in the context of regression models are poorly equipped to assess backtest overfitting.
7. Hold-outs in particular are unreliable and easy to manipulate.
- 8. The solution is not to stop backtesting. The answer to this problem is to estimate accurately the risk of overfitting.**
9. To address this concern, we have developed the CSCV framework, which derives 5 metrics to assess overfitting:
 - **Minimum Backtest Length (MBL).**
 - **Probability of Overfitting (PBO).**
 - **Out-Of-Sample Probability of Loss.**
 - **Out-Of-Sample Performance Degradation.**
 - **Backtest Stochastic Dominance.**

THANKS FOR YOUR ATTENTION!

SECTION VII

The stuff nobody reads

Bibliography (1/2)

- Bailey, D. and M. López de Prado (2012): “The Sharpe Ratio Efficient Frontier,” *Journal of Risk*, 15(2), pp. 3-44. Available at <http://ssrn.com/abstract=1821643>
- Embrechts, P., C. Klueppelberg and T. Mikosch (2003): “Modelling Extremal Events,” Springer Verlag, New York.
- Hadar, J. and W. Russell (1969): “Rules for Ordering Uncertain Prospects,” *American Economic Review*, Vol. 59, pp. 25-34.
- Hawkins, D. (2004): “The problem of overfitting,” *Journal of Chemical Information and Computer Science*, Vol. 44, pp. 1-12.
- Hirsch, Y. (1987): “Don’t Sell Stocks on Monday”, Penguin Books, 1st Edition.
- Leinweber, D. and K. Sisk (2011): “Event Driven Trading and the ‘New News’,” *Journal of Portfolio Management*, Vol. 38(1), 110-124.
- Lo, A. (2002): “The Statistics of Sharpe Ratios,” *Financial Analysts Journal*, (58)4, July/August.
- López de Prado, M. and A. Peijan (2004): “Measuring the Loss Potential of Hedge Fund Strategies,” *Journal of Alternative Investments*, Vol. 7(1), pp. 7-31. Available at <http://ssrn.com/abstract=641702>

Bibliography (2/2)

- López de Prado, M. and M. Foreman (2012): “A Mixture of Gaussians approach to Mathematical Portfolio Oversight: The EF3M algorithm,” working paper, RCC at Harvard University. Available at <http://ssrn.com/abstract=1931734>
- Resnick, S. (1987): “Extreme Values, Regular Variation and Point Processes,” Springer.
- Schorfheide, F. and K. Wolpin (2012): “On the Use of Holdout Samples for Model Selection,” American Economic Review, 102(3), pp. 477-481.
- Van Belle, G. and K. Kerr (2012): “Design and Analysis of Experiments in the Health Sciences,” John Wiley & Sons.
- Weiss, S. and C. Kulikowski (1990): “Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning and Expert Systems,” Morgan Kaufman, 1st Edition.

Bio

Marcos López de Prado is Senior Managing Director at Guggenheim Partners. He is also a Research Affiliate at Lawrence Berkeley National Laboratory's Computational Research Division (U.S. Department of Energy's Office of Science).

Before that, Marcos was Head of Quantitative Trading & Research at Hess Energy Trading Company (the trading arm of Hess Corporation, a Fortune 100 company) and Head of Global Quantitative Research at Tudor Investment Corporation. In addition to his 15+ years of trading and investment management experience at some of the largest corporations, he has received several academic appointments, including Postdoctoral Research Fellow of RCC at Harvard University and Visiting Scholar at Cornell University. Marcos earned a Ph.D. in Financial Economics (2003), a second Ph.D. in Mathematical Finance (2011) from Complutense University, is a recipient of the National Award for Excellence in Academic Performance by the Government of Spain (National Valedictorian, 1998) among other awards, and was admitted into American Mensa with a perfect test score.

Marcos is the co-inventor of four international patent applications on High Frequency Trading. He has collaborated with ~30 leading academics, resulting in some of the most read papers in Finance (SSRN), three textbooks, publications in the top Mathematical Finance journals, etc. Marcos has an Erdős #3 and an Einstein #4 according to the American Mathematical Society.

Disclaimer

- The views expressed in this document are the authors' and do not necessarily reflect those of the organizations he is affiliated with.
- No investment decision or particular course of action is recommended by this presentation.
- All Rights Reserved.

Notice:

The research contained in this presentation is the result of a continuing collaboration with

David H. Bailey, Berkeley Lab
Jonathan M. Borwein, FRSC, FAAS
Jim Zhu, Western Michigan Univ.

The full papers are available at:
<http://ssrn.com/abstract=2308659>
<http://ssrn.com/abstract=2326253>

For additional details, please visit:
<http://ssrn.com/author=434076>
www.QuantResearch.info