HOW TO SPOT BACKTEST OVERFITTING

David H. Bailey
Lawrence Berkeley National Lab (retired), and
University of California, Davis

Marcos López de Prado
Guggenheim Partners, LLC

in collaboration with
Jonathan M. Borwein, University of Newcastle, Australia
Qiji Jim Zhu, Western Michigan University

Key points

- Backtests (i.e., historical simulations of performance) are widely employed to test and operate investment strategies.
- 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 ineffective in the context of backtest evaluation.
- Under memory effects, overfitting may lead to systematic *losses*.
- Overfitting is just one example of the misuse of mathematical and statistical methods applied to finance.
- Since most published backtests do not report the number of trials involved, many are overfit.

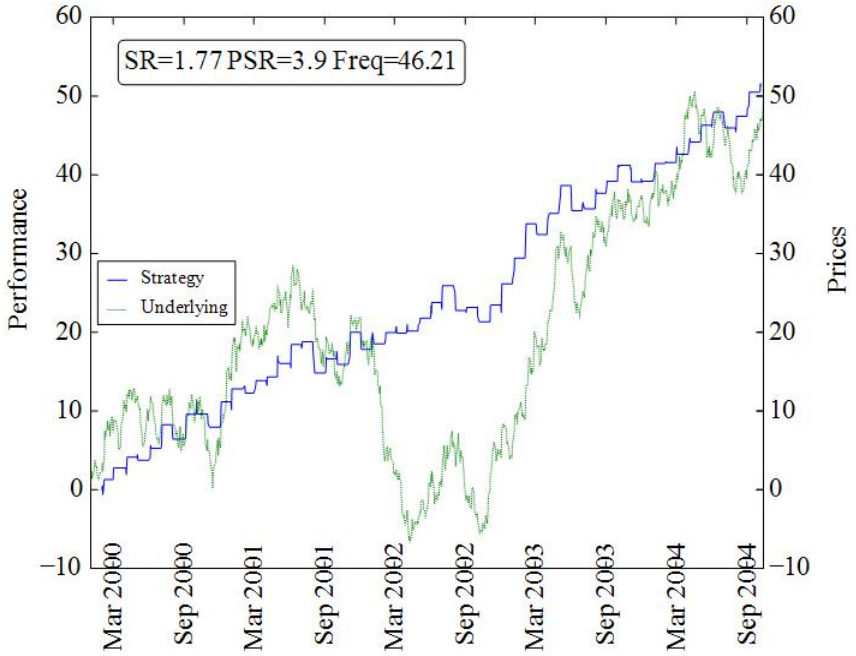
"I remember my friend Johnny von Neumann used to say, with four parameters I can fit an elephant, and with five I can make him wiggle his trunk." [Enrico Fermi, 1953]

Backtesting

- 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 a specified time period.

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.



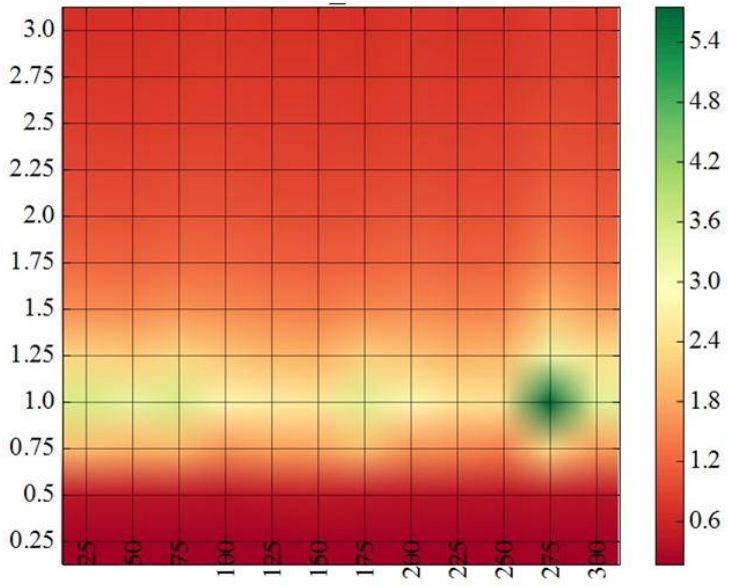
Reasons for backtesting investment strategies

- The information contained in the reported series of profits and losses may be summarized in popular performance metrics, such as the Sharpe Ratio (SR).
- These metrics are essential to select optimal parameter 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 – see graph ➔

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 in-sample Sharpe Ratio.



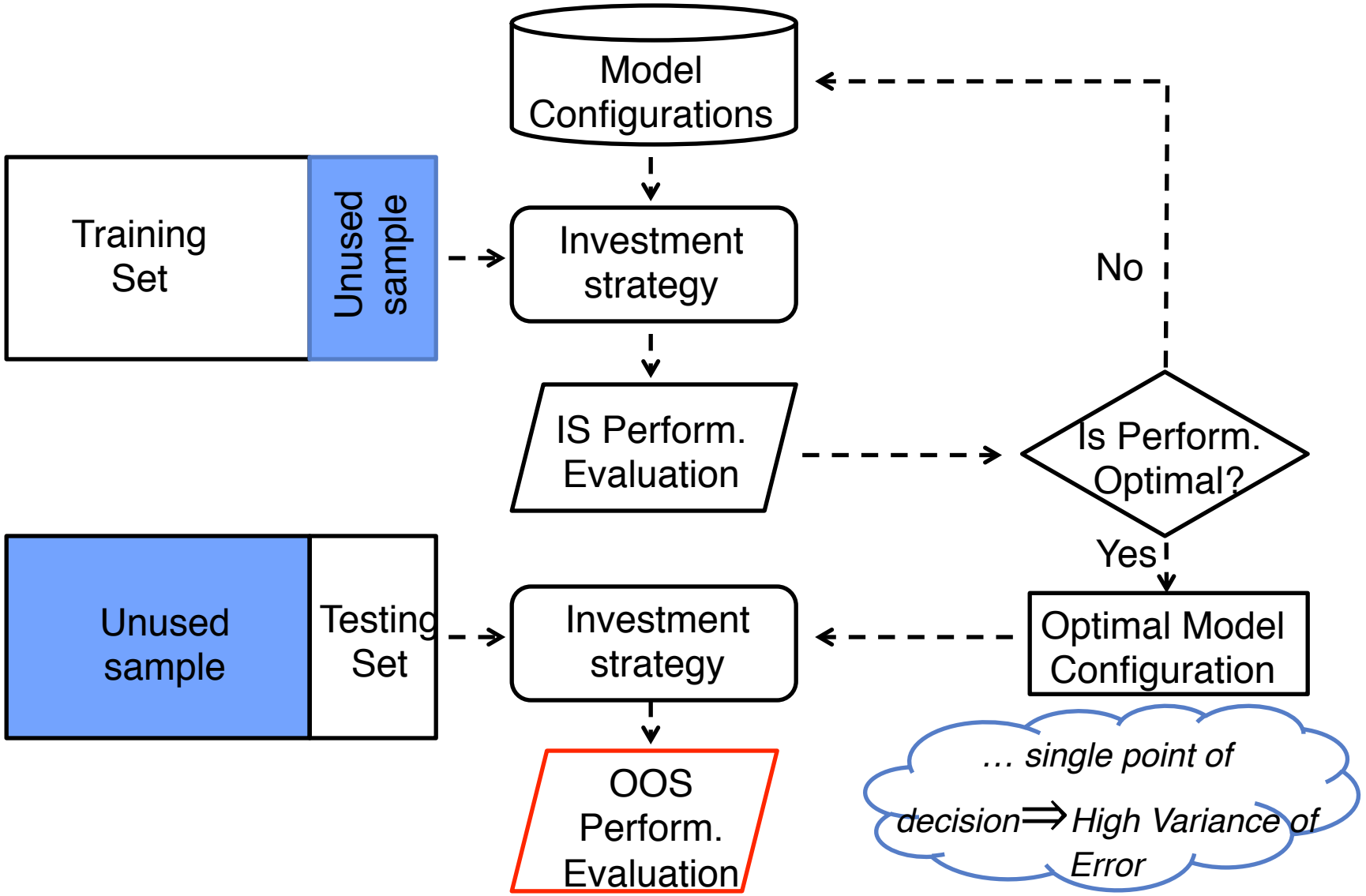
DANGER AHEAD

Supercomputers and high-tech mathematical finance algorithms can generate nonsense faster than ever before!

The principal danger is statistical overfitting of backtest data:

- When a computer can analyze thousands or millions of variations of a given strategy, it is *almost certain* that the best such strategy, measured by backtests, will be overfit (and thus of dubious value).
- Many studies claim profitable investment strategies, but their results are based only on *in-sample* (IS) statistics, with no *out-of-sample* (OOS) testing.
- Overfitting is the most common reason that mathematical investment schemes look great in backtests, but then fall flat in the real world.
- … and yet, most backtesting software does not control for the probability of backtest overfitting!

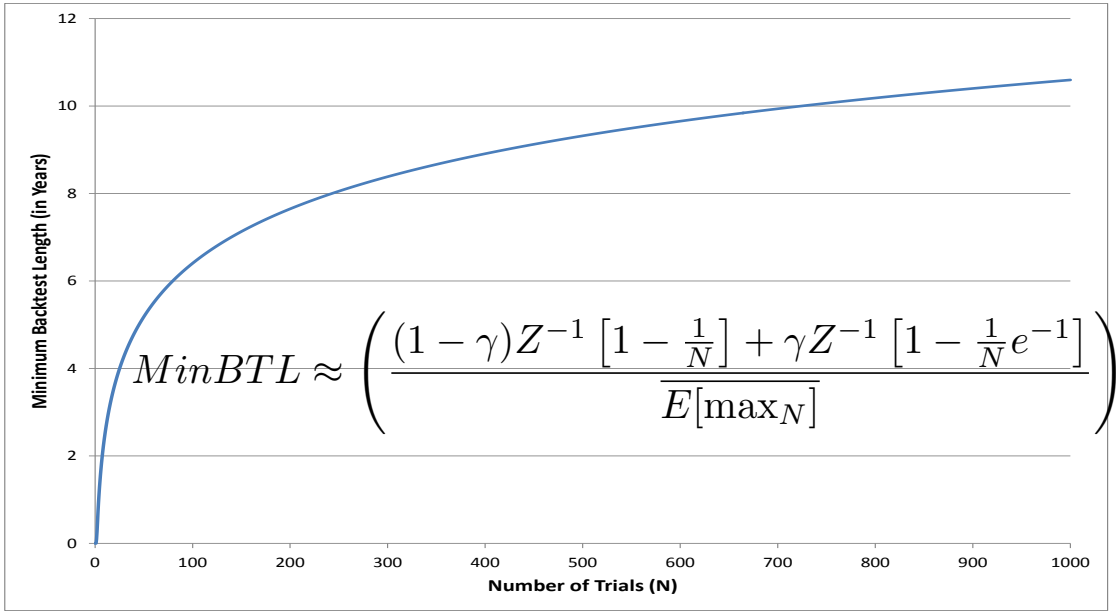
The hold-out method to test an investment strategy (not very good)



How easy is it to overfit a backtest?

Unfortunately, too easy!

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 max IS SR = 1, whereas the expected OOS SR = 0.

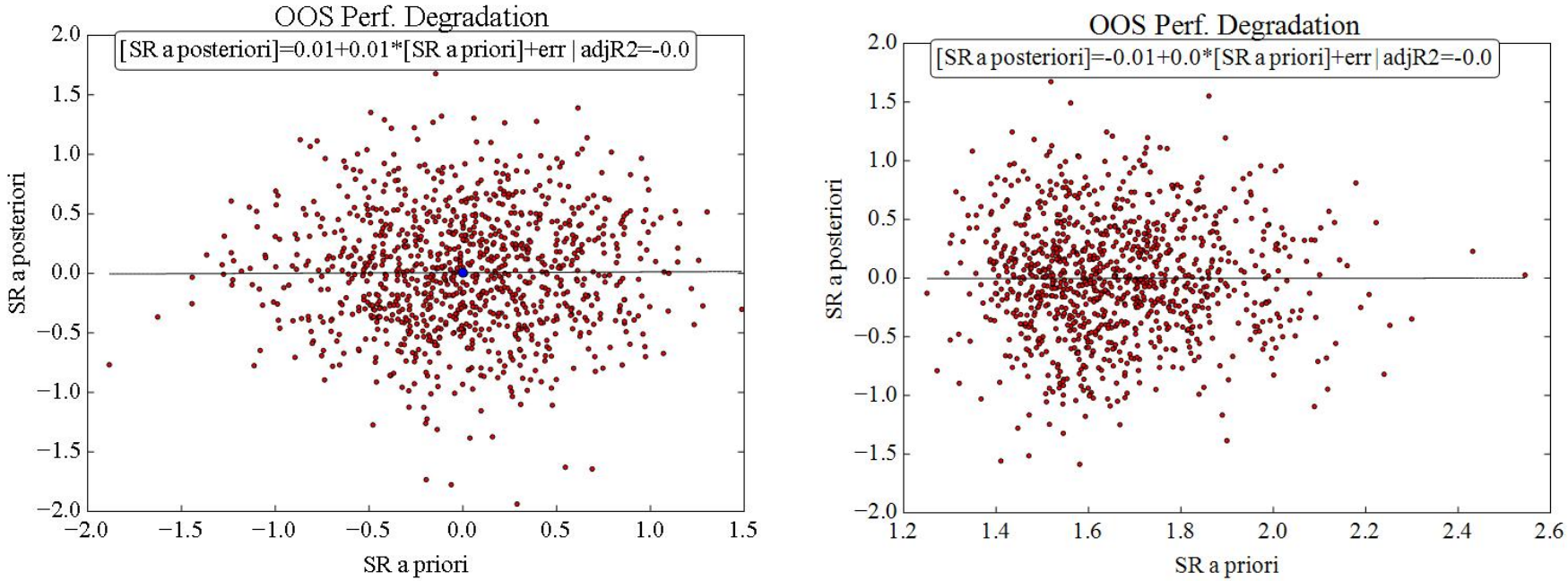


After trying only 7 independent strategy configurations, the expected maximum IS SR = 1 for a 2-year-long backtest, while the expected OOS SR = 0.

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 in-sample: The researcher just needs to keep trying alternative parameters for that strategy!

Overfitting in the absence of memory



The left figure shows the relation between SR IS (x-axis) and SR OOS (y-axis). Because the process follows a random walk, the scatter plot has a circular shape centered in the point (0,0).

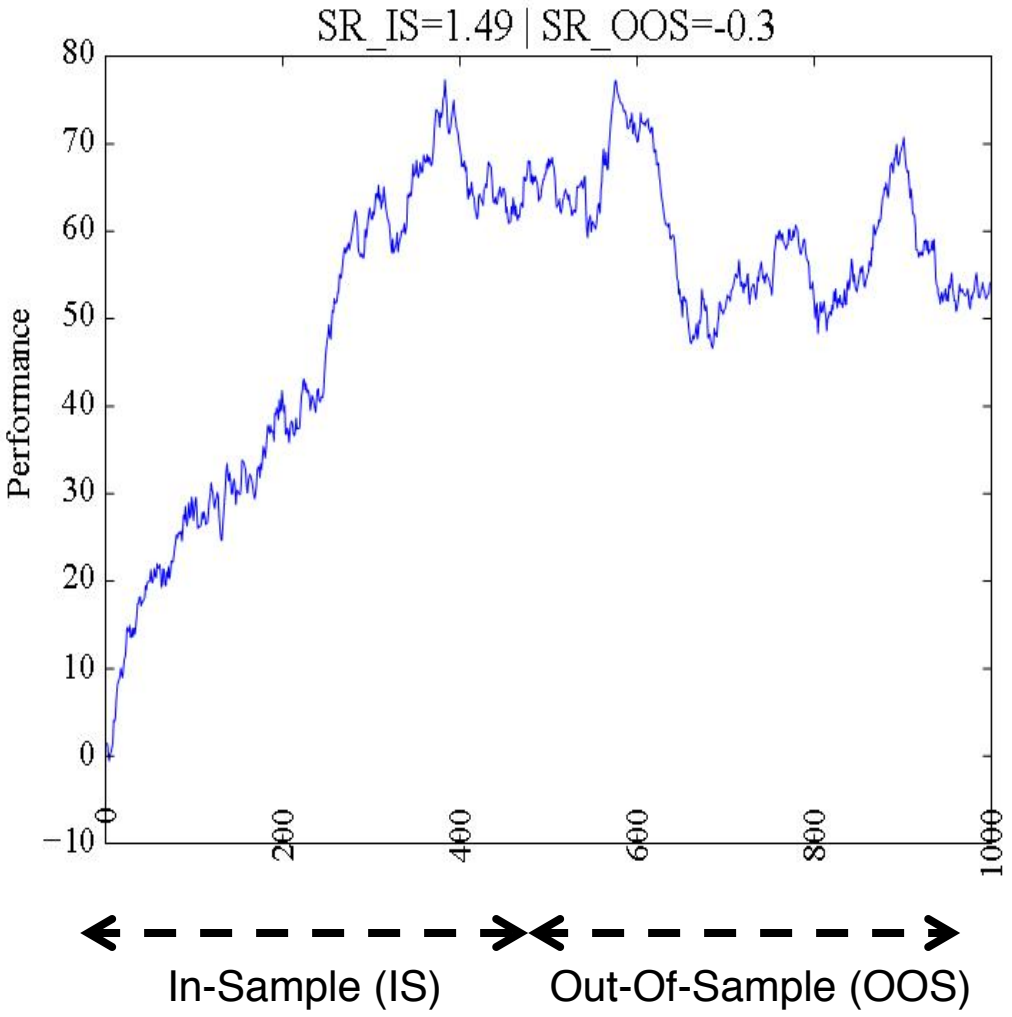
The right figure illustrates what happens once we add a “model selection” procedure. Now the SR IS ranges from 1.2 to 2.6, and it is centered around 1.7. Although the backtest for the selected model generates the expectation of a 1.7 SR, the expected SR OOS is unchanged around 0.

Overfitting in the absence of memory (cont.)

This figure shows what happens when we select the random walk with highest SR in-sample (IS).

The performance of the first half was optimized (IS), and the performance of the second half is what the investor receives out-of-sample (OOS).

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

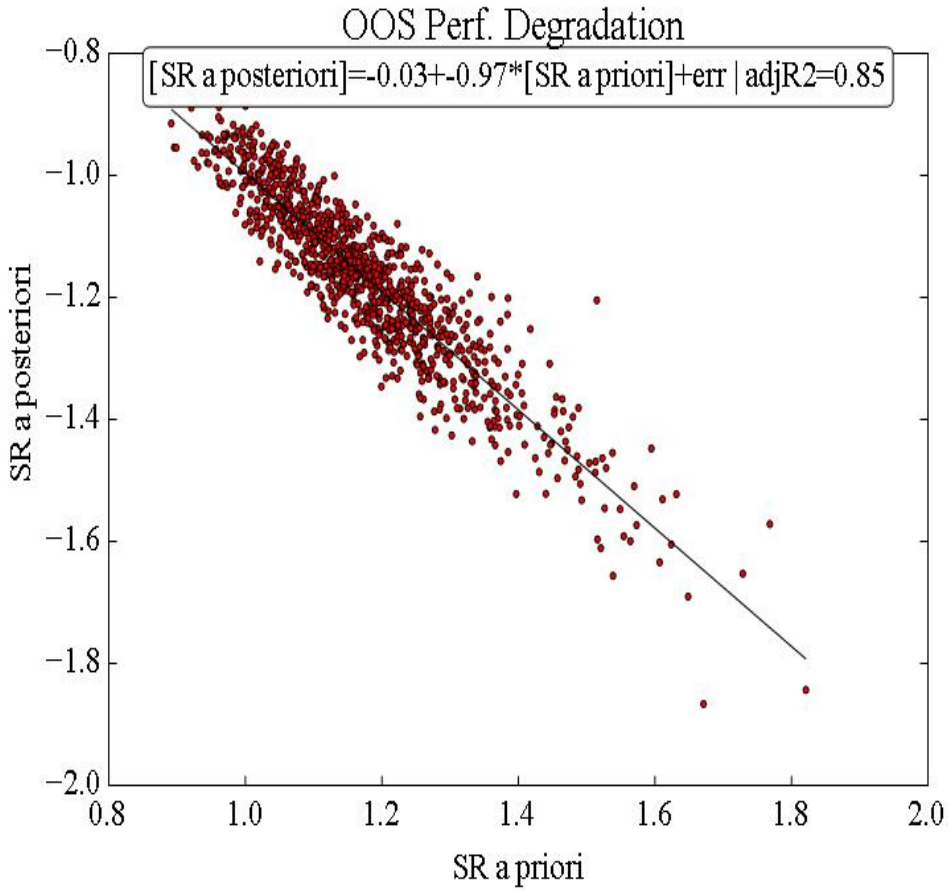


Overfitting in the presence of memory

Memory effects may cause OOS performance to be negative, even though the underlying process was trendless.

Also, a strongly negative linear relation between performance IS and OOS may arise, indicating that **the more we optimize in-sample, the worse is OOS performance**.

Conclusion:
When financial analysts do not control for overfitting, “Past performance is not an indicator of future performance” is too optimistic! Good backtest performance may be an indicator of *negative* future results.



The p-values associated with the intercept and the in-sample 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.

Tools to prevent backtest overfitting

1. Compute the probability of backtest overfitting, using a formula given in our paper "The probability of backtest overfitting," available at http://ssrn.com/abstract=2326253 or http://www.financial-math.org.
2. Compute performance degradation and probability of loss (also given in the above paper).
3. Apply the theory of stochastic dominance, which allows us to rank investment strategies without having to make assumptions regarding an individual's utility function (see above paper for details).
4. Perform model sequestration: Announce a proposed investment strategy to others (either publicly, or within a firm), then subsequently publish the results of using this strategy for a pre-specified period of time.
   - See D. Leinweber and K. Sisk, "Event Driven Trading and the 'New News'," *Journal of Portfolio Management*, vol. 38(1), pg. 110-124.

Reproducibility in finance

Using rigorous methods in mathematical finance (e.g., to prevent backtest overfitting) enhances reproducibility and reliability:

- Many other scientific disciplines are facing similar issues of reproducibility, to overcome the bias of only publishing "good" results.
- There is a growing movement in the pharmaceutical industry to require the results of all prototype drug tests to be made public. See http://www.alltrials.net.
- Johnson & Johnson recently announced it will make all test results public.
- Mathematicians and computer scientists are setting standards for reproducibility in the field of scientific computing. See:
  - V. Stodden, D. Bailey, J. Borwein, E. LeVeque, W. Rider, and W. Stein, "Setting the default to reproducible: Reproduciblity in computational and experimental mathematics," February 2013, available at http://www.davidhbailey.com/dhbpapers/icerm-report.pdf.

An absurd investment program

- An investment advisor initially sends 10,240 letters to prospective clients. In 5120 of these letters, she predicts that a certain set of securities will go up; in the other 5120 she predicts they will go down.
- One month later, if the securities have gone up, she sends another letter to the first 5120 and ignores the second 5120 (or the reverse if the securities have gone down). In 2560 of these letters, she predicts the securities will go up; in the other 2560, she predicts the securities will go down.
- One month later, if the securities have gone up, she sends another letter to the first 2560 and ignores the second 2560 (or the reverse if the securities have gone down). In 1280 of these letters, she predicts the securities will go up; in the other 1280, she predicts the securities will go down. This is repeated for ten months.
- After ten months, the remaining 10 investors, astounded by the advisor's uncanny prophetic powers to date, will entrust all their money to her.

Clearly this is an absurd, even fraudulent investment program, because investors are never told of the many other failed recommendations.

But why is backtest overfitting, where one does not disclose how many models were tested, any different?

Why the silence in the mathematical finance community?

- Historically scientists have led the way in exposing those who utilize pseudoscience to extract a commercial benefit – i.e., in the 18th century, physicists exposed the nonsense of astrologers.
- Yet financial mathematicians in the 21st century have remained disappointingly silent with those who, knowingly or not:
  - Fail to disclose the number of models that were used to develop a scheme (i.e., backtest overfitting).
  - Make vague predictions that do not permit rigorous testing and falsification.
  - Misuse charts and graphs: "Beware of fund managers bearing double y-axes." See Matthew Obrien's article on the "scary chart" in the *Atlantic* (11 Feb 2014).
  - Misuse probability theory, statistics and stochastic calculus.
  - Misuse technical jargon: "stochastic oscillators," "Fibonacci ratios," "cycles," "Elliot wave," "Golden ratio," "parabolic SAR," "pivot point," "momentum", and others in the context of finance.
- Our silence is consent, making us accomplices in these abuses.

"*One has to be aware now that mathematics can be misused and that we have to protect its good name*." – Andrew Wiles, *New York Times*, 4 Oct 2013.

Mathematicians Against Fraudulent Financial and Investment Advice (MAFFIA)

http://www.financial-math.org	(main site)
http://www.m-a-f-f-i-a.org	(alias to main site)
http://www.financial-math.org/blog/	(blog)

The principal purpose is education, not confrontation – helping readers recognize and avoid fallacies and abuses in mathematical finance.

For full technical details on the material in this talk:

- "Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample performance," *Notices of the American Mathematical Society*, to appear (May 2014), http://ssrn.com/abstract=2308659.
- "The probability of backtest overfitting," manuscript, 10 Feb 2014, available at http://ssrn.com/abstract=2326253.
- These papers (and this talk) are also available at http://www.financial-math.org.

THANK YOU!