Reproducibility and statistical overfitting in quantitative finance

David H. Bailey
http://www.davidhbailey.com
Lawrence Berkeley National Lab (retired) and Univ. of California, Davis

Collaborators: Jonathan M. Borwein (Univ. of Newcastle, Australia), Marcos Lopez de Prado (Guggenheim and LBNL), Jim Q. Zhu (Univ. of Western Michigan) and others

November 11, 2015

Reproducibility crises in physics and cosmology

- In 2011, an international team of researchers at the Gran Sasso Laboratory in Italy announced that neutrinos had exceeded the speed of light, thus directly challenging Einstein's relativity. However, after months of careful checking, a subtle flaw was found in the measurement apparatus (a major embarrassment).
- In 2013, CERN researchers confirmed the discovery of the long-sought Higgs boson. But more recently, scientists have raised questions as to whether the particle discovered is really the Higgs – it might be some other particle or particles masquerading as the Higgs; additional research studies are required.
- In March 2014, researchers announced with considerable fanfare that they had detected the fingerprint of the long-hypothesized inflationary epoch, a tiny fraction after the big bang. Sadly, within a few weeks critics pointed out that their experimental results might well be due to dust in the Milky Way, pending better data.

Reproducibility crises in biomedicine, psychology, economics, finance, autos

- In 2011, Bayer researchers reported that they were able to reproduce only 17 of 67 pharma studies.
- In 2012, Amgen researchers reported that they were able to reproduce only 6 of 53 cancer studies.
- In August 2015, the Reproduciblity Project in Virginia reported that they were able to reproduce only 39 of 100 psychology studies.
- In September 2015, the U.S. Federal Reserve was able to reproduce only 29 of 67 economics studies.
- In 2014-2015, "backtest overfitting" emerged as a major problem in computational finance.

Reproducibility Project staff

Credit: NY Times

- In March 2014, West Virginia researchers reported that they were unable to reproduce Volkswagen's claimed emission figures. This has now exploded into a major international scandal.

Big data science: DANGER AHEAD

Supercomputers operating on big data can generate nonsense faster than ever before!

Key concerns:

- Are the results statistically sound?
- Are the results numerically reliable?
- Have the results been validated using independent rigorous tests?
- Are the algorithms, data sources and processing methods well documented?

Email from DHB to a financial colleague, 10 June 2013

One thing that has always puzzled me about [the financial world] is the following sort of thing: [examples cited]. Excuse me for being “dumb,” but this sort of thing seems to me to be outright nonsense. ...

After all, the stock market, by definition, contains the consensus of all available information, including the tens of thousands of stock market analysts and economists worldwide who scour every morsel of information in the business world, and then advise the leading mutual funds and pension funds. ...

In addition, ... there are thousands more very bright mathematicians using program-trading schemes, plying every trick of time series analysis, machine learning, stealth and anti-stealth that money can buy, to wriggle every conceivable angle out of the market and beat their competitors to the punch with trades. ...

So when people like those above [say] that they “know” [where] the stock market is heading, ... or that by following their strategies, John Q Public can enjoy reliable, above-market returns, this cannot have any scientific basis. ...

So why doesn’t somebody blow this whistle on this sort of thing? Am I missing something?

Response from colleague to DHB, 17 June 2013

It is not a dumb question at all. It is a question I have struggled with and which answer makes me an unhappy man. The truth is, most people in this industry are charlatans. They do not have any particular model or theory to understand the world. They are not scientists. ...

I completely agree with your assessment. The amount of nonsense ... is incredible.

The good news is, the quants are silently taking over Wall Street, thanks to high frequency and big data. For the same reason that alchemists and astrologers fought the chemists and astronomers, the market wizards are fighting the quants. So all this ... nonsense is in part the tug of that war. An attempt of the wizards to squeeze out a few more dimes.

Email from another colleague to DHB, 5 May 2015

You have written about economics and risk assessment and so I'd like to know if you have any ideas about protecting personal wealth.

I thought of you while reading Janet Tavakoli's *Decisions: Life and Death on Wall Street*. Have you read it? I turned to the book after noting it was [promoted] by Nomi Prins, another Wall Street ex-exec like Tavakoli who's been spilling the beans about Wall Street shenanigans.

Economists like Simon Johnson, Anat Admati and Joseph Stiglitz have been writing similar stories from a broader theoretical perspective, but all-in-all, all five (and they are hardly alone) describe a rigged game.

So what to do about it at the personal level? This comes down to wondering about specific things like savings accounts, CDs, stocks, bonds and annuities, life insurance and home-ownership vs renting.

A credibility crisis in finance?

- Rightly or wrongly, many believe that the quantitative finance world was at least partly to blame for the 2007-2009 worldwide financial crisis.
- Rightly or wrongly, many individual investors believe that the financial system (high-frequency trading, "dark pools," etc.) is rigged against them.
- Many are skeptical of the countless new investment funds and financial strategies claiming to be supported by historical market data.
- Financial news is replete with dubious technical jargon: "Fibonacci ratios," "cycles," "Elliott waves," "golden ratios," "parabolic SARs," "pivot points," "technical analysis," etc.

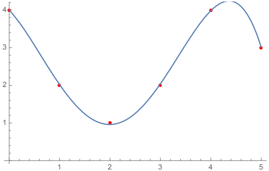
What should the financial community do?

First and foremost, the quantitative finance field needs to ensure that its own published research and commercially marketed funds and strategies are mathematically and statistically sound.

Backtest overfitting

Finance has been stung with many instances of investment strategies that look great on paper, but fall flat in practice. A leading factor is backtest overfitting:

- Proposing a model for a dataset that inherently possesses a higher level of complexity than the historical data; or
- Using a computer to try many variations of a model or strategy on the historical data, and then only presenting results from the variation that works best; or
- Constructing an exchange-traded fund by exploring millions of weighting factors, then only marketing the one with the highest backtest score.



Fitting six data points (almost perfectly!) with a fourth-degree function.

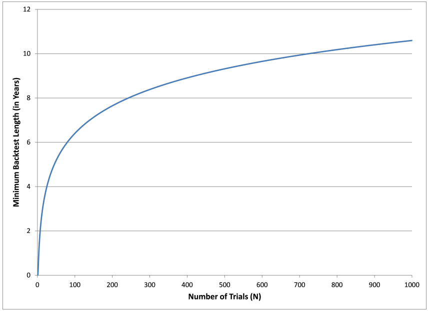
When a computer can analyze thousands, millions or even billions of variations of a given strategy on a fixed backtest dataset, it is almost certain that the best such strategy will be overfit and thus of dubious value.

How easy is it to overfit a backtest? Very!

- If only 2 years of daily backtest data are available, then no more than 7 strategy variations should be tried.
- If only 5 years of daily backtest data are available, then no more than 45 strategy variations should be tried.

A backtest that does not report the number of trials N makes it impossible to assess the risk of overfitting.

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



- "Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample performance," *Notices of the AMS*, May 2014, pg. 458–471, `http://www.financial-math.org`.

Letters to clients: An absurd investment scheme

- A financial advisor sends letters to $10,240 = 10 \times 2^{10}$ prospective clients, with 5120 predicting a certain stock will go up, and 5120 predicting it will go down.
- One month later, the advisor sends letters only to the 5120 investors who were previously sent the correct prediction, with 2560 letters predicting a certain stock will go up, and 2560 predicting it will go down.
- After ten months, the final ten investors will have been sent ten consecutive spot-on predictions!

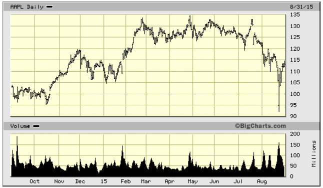
This strategy is absurd, even fraudulent, because the final ten investors are not told of the thousands of other letters with different predictions.

But why is marketing a statistically overfit strategy, where potential investors are not informed of the millions of failed computer trials behind the strategy, any different?

A not-so-absurd investment strategy

Suppose an investor believes that there are daily, weekly or monthly patterns in stock market data, and she seeks to exploit them. Sample strategies:

- Basic strategy: Buy a set of stocks each Monday, then sell on Wednesday; buy on the 6th of the month, then sell on the 19th; sell in May and go away, etc.
- Refinements: Sell the portfolio if it drops more than 10% from start; purchase shares only when they increase in value more than 10% from start; etc.



Apple stock price

31 Aug 2014 – 31 Aug 2015

Even with these very simple strategies, there are literally millions of variations (by changing various parameters), which can be quickly explored by computer.

Selecting only the best combination of parameters (and not mentioning the many others that were tried) is a classic selection bias statistical error.

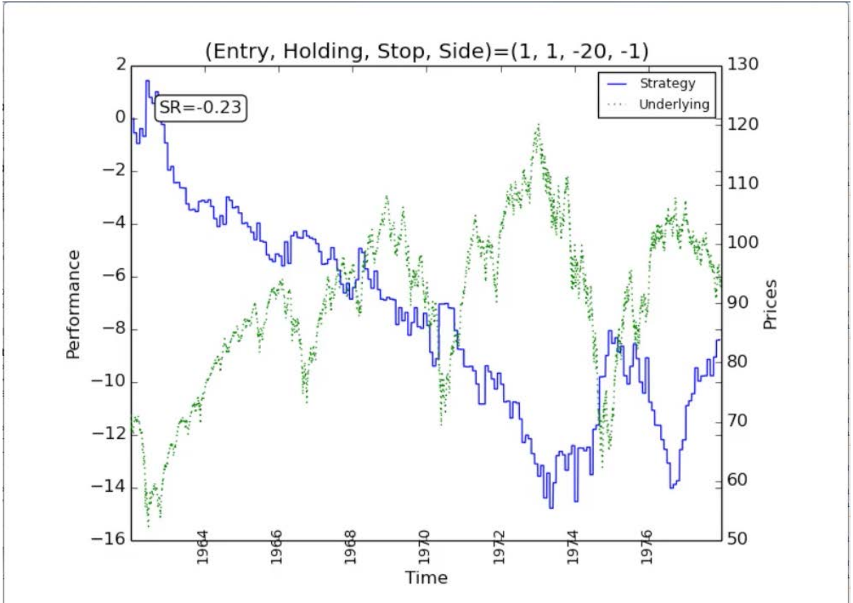
Backtest overfitting: An interactive example

An online demonstration is backtest overfitting is now available:

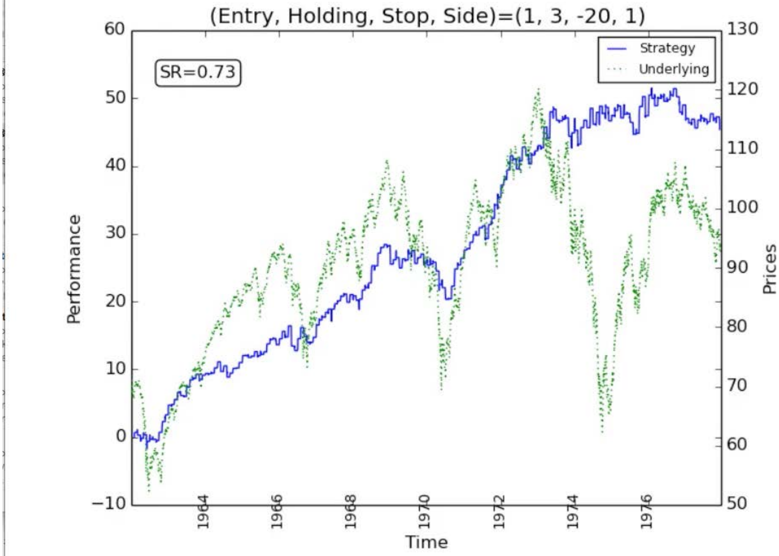
- The user can select either pseudorandom data or real S&P500 historical data.
- The program then runs a simple monthly-cycle strategy with parameters (day in, holding period, stop-loss percentage, side, etc.), adjusting the parameters to find an optimal strategy.
- The final optimal strategy is then tried on a new (out-of-sample) dataset.
- This software is now available in an online demo (try it yourself!): `http://www.financial-math.org`

- Credits: Stephanie Ger, Marcos Lopez de Prado, Amir Salehipour, Alex Sim and Kesheng Wu.

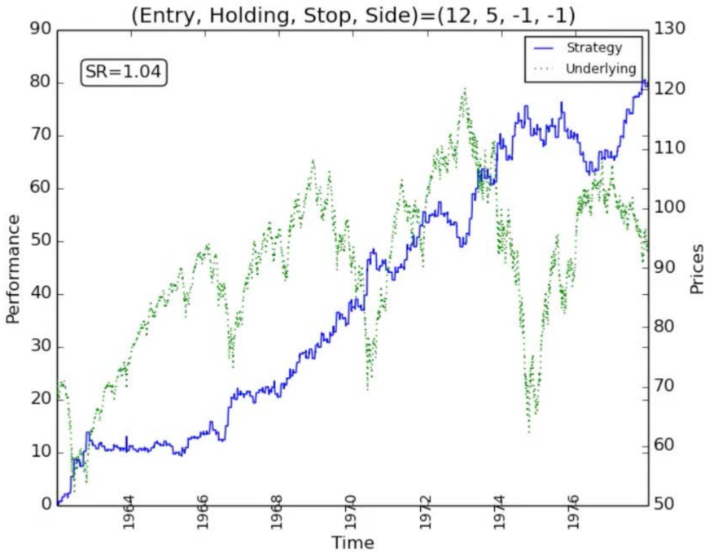
Initial strategy on input data (S&P500, 1960–1980): Sharpe ratio = -0.23



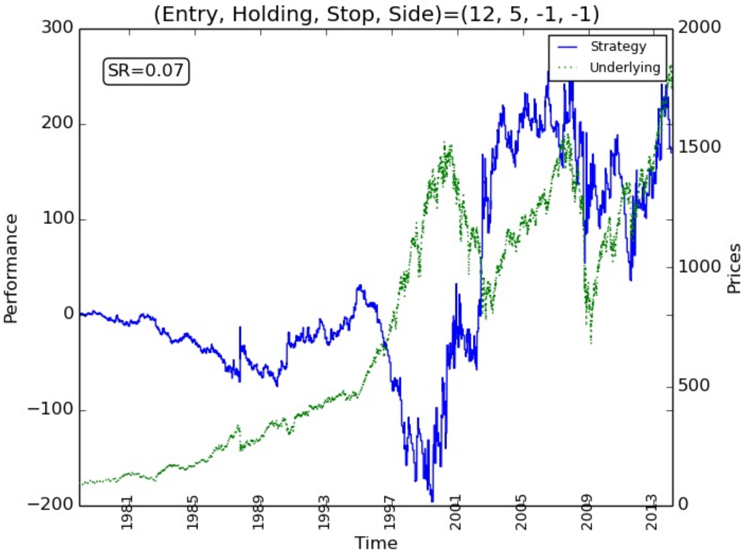
Improved strategy: Sharpe ratio = 0.73



Final (optimal) strategy: Sharpe ratio = 1.04



Final strategy on new data (S&P500, 1980–2013): Sharpe ratio = 0.07



Analysis

- After exploring the very large space of strategy variations, the computer program found a strategy that achieved a Sharpe ratio of 1.04 on the input (backtest) data (S&P500, 1960–1980).
- However, this optimal strategy, when applied to new (out-of-sample) data (S&P500, 1980–2013), failed miserably — the Sharpe ratio was 0.07.
- In other words, the "optimal" strategy found by the computer search only fit idiosyncrasies of the input (backtest) dataset — it has no "intelligence."

The software demo program is now available online:
http://www.financial-math.org

For additional analysis, aimed at a fairly basic level, see:

- D. H. Bailey, S. Ger, M. Lopez de Prado, A. Sim and K. Wu, "Statistical overfitting and backtest performance," manuscript, 07 Oct 2014, http://www.financial-math.org.

New papers on backtest overfitting by DHB and colleagues

- Presents formulas relating size of dataset to likelihood of backtest overfitting:
  D. H. Bailey, J. M. Borwein, M. Lopez de Prado and Q. J. Zhu, “Pseudo-mathematics and financial charlatanism: The effects of backtest over fitting on out-of-sample performance,” *Notices of the American Mathematical Society*, May 2014, pg. 458–471.
- Presents formulas for calculating the probability of backtest overfitting:
  D. H. Bailey, J. M. Borwein, M. Lopez de Prado and Q. J. Zhu, “The probability of backtest overfitting,” *Journal of Computational Finance*, to appear, 27 Feb 2015.
- Introduces backtest overfitting for a general audience:
  D. H. Bailey, S. Ger, M. Lopez de Prado, A. Sim and K. Wu, “Statistical overfitting and backtest performance,” manuscript, 07 Oct 2014.
- Defines a “deflated Sharpe ratio,” correcting for some forms of distortion:
  D. H. Bailey and M. Lopez de Prado, “The deflated Sharpe ratio: Correcting for selection bias, backtest overfitting and non-normality,” *Journal of Portfolio Management*, to appear, 31 Jul 2014.

Full references and preprints are available at: `http://www.financial-math.org`

Why the silence in the mathematical finance community?

Historically scientists have exposed those who utilize pseudoscience for commercial gain, e.g., in the 18th century, physicists exposed the nonsense of astrologers.

Yet financial mathematicians in the 21st century have remained disappointingly silent with the regards to those in the community who, knowingly or not:

1. Fail to disclose the number of models or variations that were used to develop an investment strategy.
2. Make vague predictions that do not permit rigorous testing and falsification.
3. Misuse probability theory, statistics and stochastic calculus.
4. Use pseudomathematical jargon: "Fibonacci ratios," "cycles," "Elliott waves," "golden ratios," "parabolic SARs," "pivot points," "technical analysis," etc.

As we wrote in a recent paper:

"Our silence is consent, making us accomplices in these abuses."

- "Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample performance," *Notices of the AMS*, May 2014, pg. 458–471, `http://www.financial-math.org`.

One financial colleague's recommendation

Empirical finance is in crisis: Our most important discovery tool is historical simulation, and yet, most backtests and time series analyses published in journals are flawed. The problem is well-known to professional organizations of statisticians and mathematicians, who have publicly criticized the misuse of mathematical tools among finance researchers. ...

In an attempt to overcome the challenges posed by multiple testing and selection bias, I emphasize the need to move from an individual-centric to a community-driven research paradigm. ... Stronger theoretical foundations and closer ties [between academics and] financial firms would help prevent false discoveries.

- M. Lopez de Prado, "The future of empirical finance," *Journal of Portfolio Management*, 41(4), 2015, `http://www.financial-math.org`.

Visit our website and blog

Mathematicians Against Fraudulent Financial and Investment Advice (MAFFIA):
http://www.financial-math.org
or
http://www.m-a-f-f-i-a.org

This talk is available at:
http://www.davidhbailey.com/dhbtalks/dhb-fin-repro.pdf