Stock portfolio design and backtest overfitting

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

Collaborators: Marcos Lopez de Prado (Guggenheim Partners and LBNL), Jim Q. Zhu (Univ. of Western Michigan) and others

November 3, 2016

Reproducibility crises in biomedicine, psychology, economics, finance

- 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 2015, "backtest overfitting" emerged as a major problem in computational finance.

Reproducibility Project staff

Credit: NY Times

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.

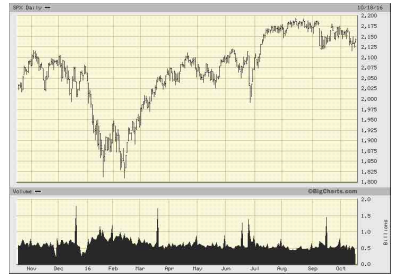
I thought of you while reading Janet Tavakoli's *Decisions: Life and Death on Wall Street*. Have you read it? I turned to the book after noting it was [promoted] by Nomi Prins, another Wall Street ex-exec like Tavakoli who's been spilling the beans about Wall Street shenanigans.

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 hundreds of new investment funds and strategies that are produced each year.
- Financial news is replete with pseudomathematical charts and technical jargon: "Fibonacci ratios," "cycles," "Elliott waves," "golden ratios," "parabolic SARs," "technical analysis," "pivot points," "symmetrical triangles," "rising wedges," etc.

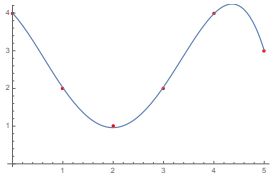


What should the mathematical finance community do? Ensure that its own published research and strategies are mathematically and statistically sound.

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 factor sets, then only marketing the one with the highest backtest score.



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

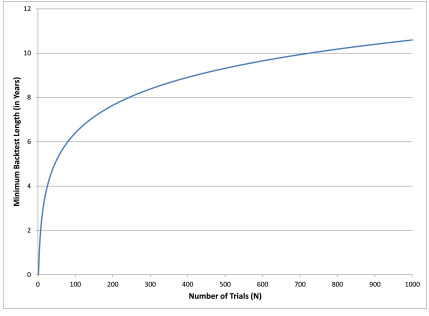
When a computer can analyze thousands, millions or even billions of variations of a fund or strategy on a fixed backtest dataset, it is almost certain that the optimal fund or strategy will be overfit and thus of dubious value.

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, 458–471.

Letters to clients: An absurd investment scheme

- A financial advisor sends letters to $5,120 = 5 \times 2^{10}$ prospective clients, with 2560 predicting a certain stock will go up, and 2560 predicting it will go down.
- One month later, the advisor sends letters only to the 2560 investors who were previously sent the correct prediction, with 1280 letters predicting a certain stock will go up, and 1280 predicting it will go down.
- After ten months, the final five investors will have been sent ten consecutive spot-on predictions!

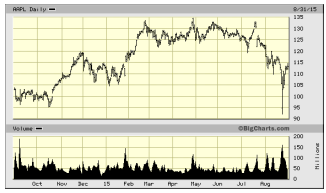
This strategy is absurd, even fraudulent, because the final five investors are not told of the 10,235 other letters with different predictions.

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.

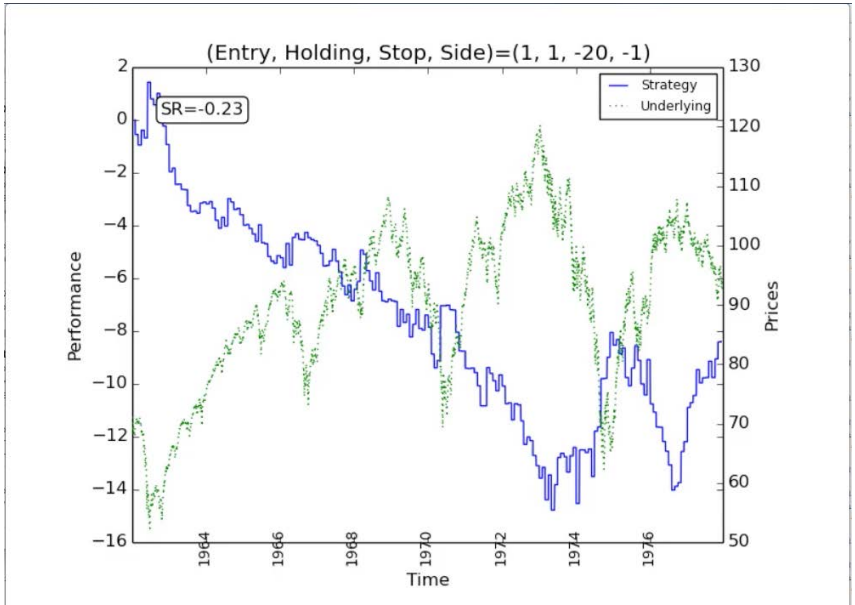
Backtest overfitting: An interactive example

An online demonstration is backtest overfitting is now available:

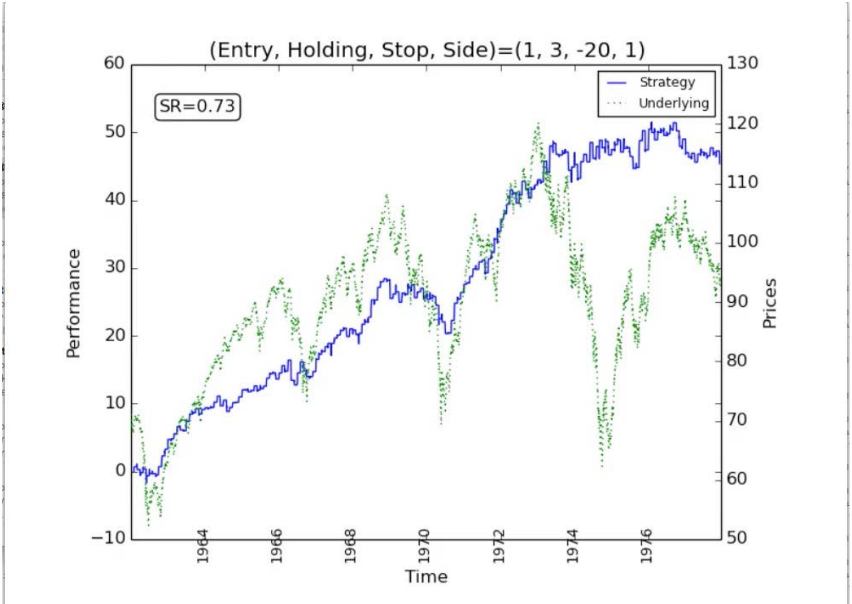
- The user can select either pseudorandom data or real S&P500 historical data.
- The program then runs a simple monthly-cycle strategy with parameters (day in, holding period, stop-loss percentage, side, etc.), adjusting the parameters to find an optimal strategy.
- The final optimal strategy is then 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.

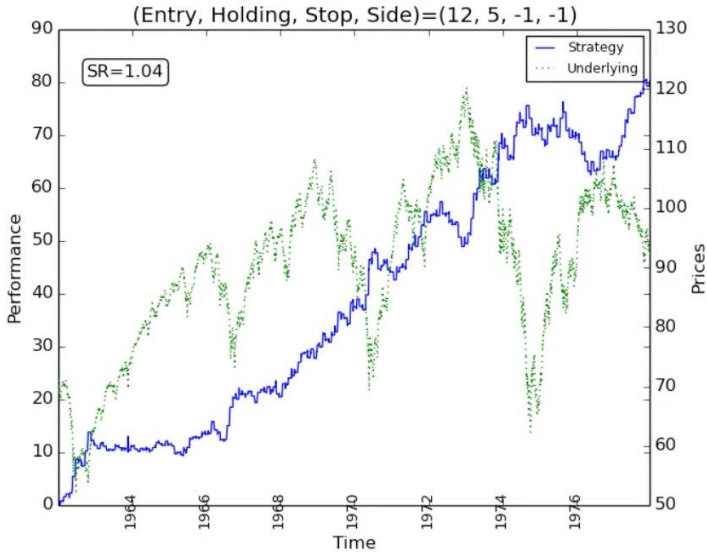
Initial strategy on input data (S&P500, 1960–1980): Sharpe ratio = -0.23



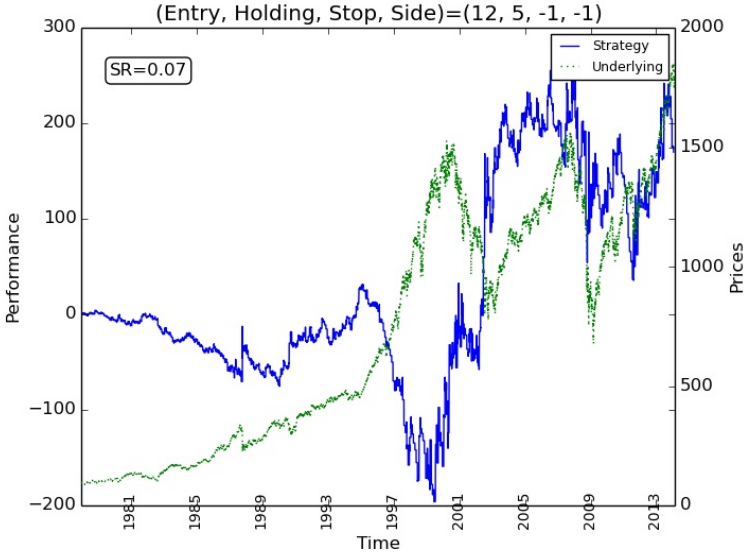
Improved strategy on input data: Sharpe ratio = 0.73



Final (optimal) strategy on input data: Sharpe ratio = 1.04



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



Additional details on backtest overfitting

- 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, 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, 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, 2014.

Preprint copies are available at: `http://www.financial-math.org`

Proliferation of new stock funds

- Roughly USD$2.1 trillion is held in U.S.-listed exchange-traded funds (ETFs), with hundreds minted each year.
- In a 2012 study, researchers found that the median time between the definition of a new index and the inception of a new exchange-traded fund based on the index dropped from almost three years in 2000 to only 77 days in 2011.
- As a result, the report concludes, "most indexes have little live performance history for investors to assess in the context of a new ETF investment."
- Out of 370 new indexes, 87% of the indexes outperformed the broad U.S. stock market over the time period used for the backtest, but only 51% outperformed the broad market after inception of the index.
- In particular, the study found an average 12.25% annualized excess return above the broad U.S. stock market for a five-year backtest, but -0.26% excess return in the five years following the inception of the index.

https://pressroom.vanguard.com/content/nonindexed/7.23.2012_Joined_at_the_hip.pdf

Designing a stock fund portfolio to match a desired performance profile

- Given some desired performance profile (a time series), we construct a weighted subset of S&P500 stocks whose performance matches, as closely as possible, that of the profile over the specified backtest time period.
- The design minimizes the sum of squares deviation of the weighted portfolio time series from the given profile time series.
- In a typical run, some of the resulting weights are negative, corresponding to shorted positions in certain stocks. This potentially exposes the portfolio to losses.
- As an alternate option, the weights are calculated subject to the constraint that all weights must be greater than or equal to zero.

- D. H. Bailey, J. M. Borwein and M. Lopez de Prado, "Stock portfolio design and backtest overfitting," to appear in *Journal of Investment Management*, preprint available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2739335.

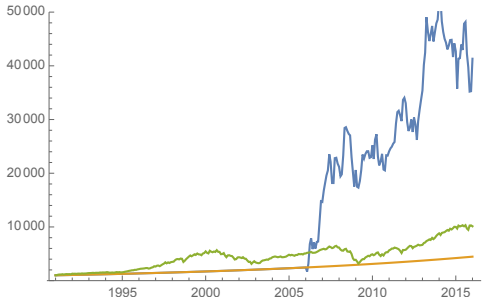
Three types of performance profiles

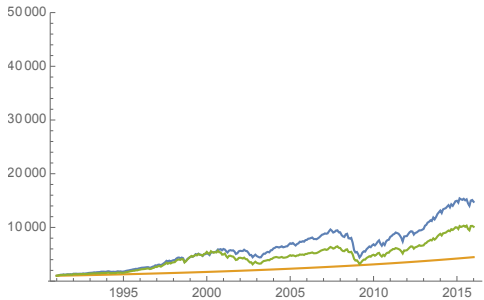
Using our program, one can generate any of several target profiles, including (here p is an annual percentage rate):

1. *Steady capital growth*: A steady increase by the fraction $(1 + p/(100r))$ per time period (i.e., growing by p/r percent each time period, where r is the number of time periods per year; e.g., $r = 12$).
2. *Stair-step growth*: A stair-step function that is constant, except that at the end of each q-year period it increases by the fraction $(1 + p/(100r))^{qr}$ (i.e., at the end of each q-year period, it increases by a full q years' growth of Profile 1 above). We took $q = 1$ in the examples below.
3. *Sinusoidal growth*: A sinusoidal function that increases by the fraction $(1 + p/(100r))$ per time period, as in profile #1, but is multiplied by a sine wave that varies from $1/2$ to $3/2$, with period q years. We took $q = 5$.

The second and third profiles are included mainly to illustrate that *any* function whatsoever may be specified for the profile.

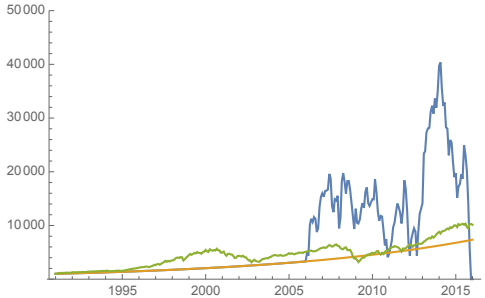
Steady growth profile, APR = 6%

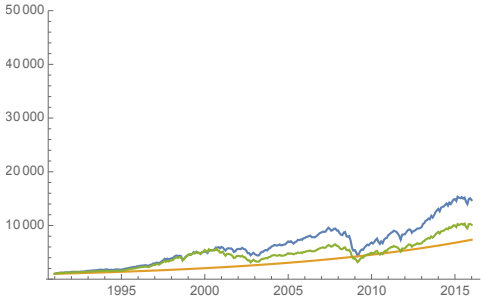




Standard portfolio (L) and all-positive portfolio (R). Blue: portfolio; orange: target profile; green: S&P500.

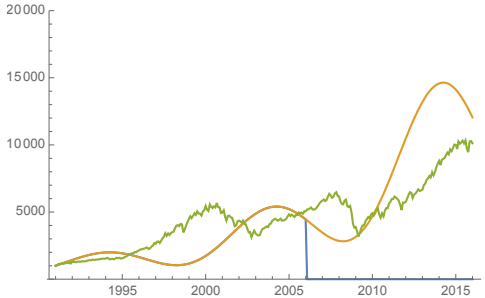
Steady growth profile, APR = 8%

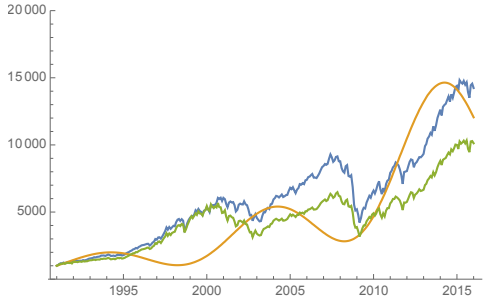




Standard portfolio (L) and all-positive portfolio (R). Blue: portfolio; orange: target profile; green: S&P500.

Sinusoidal profile, APR = 10%





Standard portfolio (L) and all-positive portfolio (R). Blue: portfolio; orange: target profile; green: S&P500.

Summary of 20 runs

			Standard weights				All-positive weights			
			RMS dev.		Sharpe ratio		RMS dev.		Sharpe ratio	
Profile	Fig.	APR	IS	OOS	IS	OOS	IS	OOS	IS	OOS
Steady growth	1	6%	0.000	7.658	-0.120	0.168	1.426	1.910	0.163	-0.025
	2	8%	0.000	2.534	-0.079	FAIL	1.016	0.970	0.162	-0.025
	3	10%	0.000	0.996	-0.038	FAIL	0.695	0.391	0.161	-0.026
	4	12%	0.000	1.178	0.003	FAIL	0.452	0.276	0.157	-0.027
	5	15%	0.000	5.953	0.065	0.178	0.223	0.557	0.145	-0.016
	6	18%	0.000	0.996	0.126	FAIL	0.218	0.711	0.177	-0.021
Stair-step	7	8%	0.000	9.395	-0.066	0.167	1.086	1.039	0.162	-0.025
	8	10%	0.000	0.996	-0.024	FAIL	0.768	0.442	0.161	-0.025
Sinu-soidal	9	8%	0.000	4.518	-0.064	FAIL	1.584	1.528	0.162	-0.024
	10	10%	0.000	0.996	-0.029	FAIL	1.267	0.867	0.158	-0.024

"APR": annual percentage rate; "IS": in-sample period, 1991–2005 (15 years); "OOS": out-of-sample period, 2006–2015 (10 years); "RMS dev.": root-mean-square deviation from target profile; "Sharpe ratio": Sharpe ratio relative to S&P 500 with reinvested dividends; "FAIL": 100% loss of capital.

Analysis

- In every case, the standard portfolio performance achieved zero deviation over the in-sample period. Only beginning with 2006 (the out-of-sample period) do the blue curves depart from the orange curves in the plots.
- In some cases the standard portfolios did remarkably well, but in other cases they failed catastrophically.
- The central objective here, namely to achieve, by means of a weighted portfolio of S&P 500 stocks, a desired performance profile that also holds on *out-of-sample data*, is certainly not met.
- The positive-weight portfolios are significantly less erratic and often outperform both the target profile and the S&P 500 benchmark). But these portfolios fail to match the target profiles either in-sample or out-of-sample.

"Beating the market" and backtest overfitting

- Overfitting, and erratic performance, is unavoidable in this or any scheme that amounts to searching over a large set of strategies or fund weightings, and only implementing or reporting the final optimal scheme.
- The same difficulty afflicts many other attempts to construct an investment strategy based solely on daily, weekly, monthly or yearly historical market data, such as with charts (as is often done by technical analysts) or tracking a particular risk profile, as many smart beta ETFs attempt.
- By and large, any underlying actionable information that might exist in such data has long been mined by highly sophisticated computerized algorithms operated by large quantitative funds and other organizations.
- Any lesser efforts, such as those described here, are doomed to be statistically overfit, and if followed may well have disastrous consequences.

Why the silence in the mathematical finance community?

Historically scientists have exposed those who utilize pseudoscience for commercial gain. 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 charts and jargon: "Fibonacci ratios," "cycles," "Elliott waves," "golden ratios," "parabolic SARs," "technical analysis," "pivot points," "symmetrical triangles," "rising wedges," etc.

As we wrote in a recent paper:
"Our silence is consent, making us accomplices in these abuses."

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.

Thanks! Visit our website at:
Mathematicians Against Fraudulent Financial and Investment Advice (MAFFIA):
http://www.financial-math.org http://www.m-a-f-f-i-a.org

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