JOIM
www.joim.com

Journal Of Investment Management, Vol. 15, No. 1, (2017), pp. 75–87

STOCK PORTFOLIO DESIGN AND BACKTEST OVERFITTING

David H. Bailey[a], *Jonathan M. Borwein*[b] *and Marcos López de Prado*[c]

In mathematical finance, backtest overfitting connotes the usage of historical market data to develop an investment strategy, where too many variations of the strategy are tried, relative to the amount of data available. Backtest overfitting is now thought to be a primary reason why investment models and strategies that look good on paper often disappoint in practice. Models and strategies suffering from overfitting typically target the specific idiosyncrasies of a limited dataset, rather than any general behavior, and, as a result, often perform erratically when presented with new data.

In this study, we address overfitting in the context of designing a mutual fund or investment portfolio as a weighted collection of stocks. Very often a newly minted equity-based fund of this type has been designed by an exhaustive computer-based search of some sort to obtain an optimal weighting that exhibits excellent performance based, say, on the past 10 or 20 years' historical market data, and the fund often highlights this backtest performance.

In the present paper, we illustrate why this backtest-driven portfolio design process often fails to deliver real-world performance. We have developed a computer program that, given any desired performance profile, designs a portfolio consisting of common securities, such as the constituents of the S&P 500 index, that achieves the desired profile based on in-sample backtest data. We then show that these portfolios typically perform erratically on more recent, out-of-sample data. This is symptomatic of statistical overfitting. Less erratic results can be obtained by restricting the portfolio to only positive-weight components, but then the results are quite unlike the target profile on both in-sample and out-of-sample data.

[a]Lawrence Berkeley National Laboratory (retired), 1 Cyclotron Road, Berkeley, CA 94720, USA, and University of California, Davis, Department of Computer Science. E-mail: david@davidhbailey.com.
[b]CARMA, University of Newcastle, NSW 2303, Australia. Borwein passed away on August 2, 2016.
[c]Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA. E-mail: lopezdeprado@lbl.gov.

1 Introduction

In mathematical finance, *backtest overfitting* connotes the usage of historical market data to develop an investment strategy, where too many variations of the strategy are tried, relative to the amount of data available. Backtest overfitting is now thought to be a primary reason why investment models and strategies that look good on paper often disappoint in practice. Models and strategies suffering from overfitting typically target the specific idiosyncrasies of a limited dataset, rather than any general behavior, and, as a result, often perform erratically when presented with new data.

Backtest overfitting is an instance of the more general phenomenon of multiple testing in scientific research, where numerous variations of a model are tested on the same data, without accounting for the potential increase in false positive rates. In finance it is common to conduct millions, if not billions, of tests on the same data in search of an optimal strategy. Since financial academic journals typically do not report the number of experiments involved in a discovery, it is possible that many published investment results are false positives.

In one earlier study (Bailey *et al.*, 2014), the present authors and Jim Zhu demonstrated that overfitting is surprisingly likely to occur in the construction of financial models and strategies. We showed, for example, that if only five years of daily backtest data are available, then no more than 45 strategy variations should be tried. In another earlier study (Bailey *et al.*, 2015), the present authors, Amir Salehipour and Jim Zhu described online tools that permit one to generate in-sample test data, based on either pseudorandom data or else real S&P 500 data, and then, by utilizing a computer search scheme, produce an "optimal" strategy. The tools then run these "optimal" strategies on out-of-sample data and evaluate their performance. In most cases, the Sharpe ratio (Sharpe, 1994) on out-of-sample data is either negative, or is significantly lower than the performance in-sample. These tools are available at http://datagrid.lbl.gov/backtest/ and https://carma.newcastle.edu.au/tenuremaker/, respectively.

2 Designing a portfolio to match a given profile

In this study, we address overfitting in the context of designing a mutual fund or investment portfolio as a weighted collection of stocks. Most funds and portfolios employ a relatively simple and objectively constructed weighting, such as capitalization weighting (employed, for example, by funds that track the standard S&P 500 index or the London FTSE 100 index) and equal weighting (employed, for example, by funds that track the S&P 500 equal-weight index).

The most rapidly growing sector of the mutual fund market is the sector of exchange-traded funds, which combine the market exposure of a mutual fund with the convenience of real-time stock trading. As of 30 June 2015, some USD$2.1 trillion was held in U.S.-listed exchange-traded funds (Vanguard Group, 2015). Hundreds of new exchange-traded funds are minted each year. There is concern in the field that some of these funds are not really independent of their indexes, and, in any event, most are not significantly different from the broad market. In a 2012 study, researchers at the Vanguard Group 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" (Vanguard Group, 2012).

How do these newly-minted exchange-traded funds perform? The 2012 Vanguard report found that out of 370 indexes for which they were able to obtain reliable information, 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, (Vanguard Group, 2012).

In the present paper, we illustrate why this backtest-driven portfolio design process often fails to deliver real-world performance. We have developed a computer program that, given *any* desired performance profile, designs a portfolio consisting of common securities, such as the constituents of the S&P 500 index, that achieves the desired profile based on in-sample backtest data. We then show that these portfolios typically perform erratically on more recent, out-of-sample data. This is symptomatic of statistical overfitting. Less erratic results can be obtained by restricting the portfolio to only positive-weight components, but then the results are quite unlike the target profile on both in-sample and out-of-sample data.

3 Mathematical techniques

In this study we design a portfolio of stocks to track an arbitrary performance profile by minimizing a simple objective function, namely the square root of the tracking errors. This objective function is also minimized when building minimum variance or minimum tracking error portfolios, hence our conclusions can be extended to applications of Markowitz-style optimization, such as the Critical Line algorithm (CLA). CLA delivers a result that is mathematically correct in-sample; however it may not be optimal out-of-sample. For a detailed discussion of CLA suboptimality out-of-sample, we refer the reader to (Bailey and López de Prado, 2013; López de Prado, 2016).

The basic approach employed by our computer program is as follows. Given a set of stocks and a desired performance profile, we employ techniques of optimization theory (Chong and Zak, 2013) to find a set of weights that minimize the sum of squares deviation of the weighted portfolio time series from the target profile time series. The resulting mathematical formulation is in the form of a matrix equation, which can be solved using widely available linear algebra software. Mathematical details are given in Appendix 1.

As we shall see below, when this technique is implemented on real stock market data, some of the resulting weights are typically negative. This means that the corresponding stocks will be shorted in the resulting portfolio. While shorting is certainly a legitimate trading strategy, shorting exposes the portfolio to potentially large losses; indeed, this occurred in several examples below. So one can also ask for an optimal set of weights subject to the constraint that each weight must be greater than or equal to zero.

Problems of this type are known as optimization problems with interval or inequality constraints. These problems have been studied at length, and quite a few sophisticated schemes have been developed for their solution (Bailey and López de Prado, 2013; Barzilai and Borwein, 1998; Bertsekas, 1982; Borwein and Lewis, 1992). In this particular application, we have employed what is known as the logarithmic barrier scheme (Gockenbach, 2003), which is to append a logarithmic term with a constant multiplier to the minimization problem. The presence of this logarithmic term penalizes very small weights and thus serves as a barrier, keeping the weights away from zero or negative values. This is not the same as solving the constrained problem, but by

successively reducing the constant multiplier, the desired limiting solution can be obtained. Mathematical details, and a statement of the resulting algorithm, are given in Appendix 2.

4 Implementation using S&P 500 stocks

We now discuss a computer implementation of these schemes. For simplicity, we base this analysis on historical stock data of stocks currently listed in the S&P 500 index, the world's most widely traded stock index. Data for S&P 500 stocks are easy to obtain online. For example, Apple Computer's daily stock closings going back to 1980 can be downloaded from https://finance.yahoo.com/q/hp?s=AAPL or http://www.google.com/finance?q=AAPL.

Our program operates as follows. Given a starting year, plus the number of years for the in-sample (L_1) and out-of-sample (L_2) tests, as well as the number r of time periods per year, the program extracts the relevant data from the database for as many stocks as possible, in order of capitalization weighting in the S&P 500 index as of 22 January 2016. Only those stocks for which a full set of data covering the time period in question are incorporated for analysis. We took monthly intervals for the analyses described in Section 5 (i.e., $r = 12$), but our program can handle monthly, quarterly or annual intervals.

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).
(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 $q = 5$.

Arguably Profile 1 is a highly desirable profile—steady capital growth, month after month, even in times of high market volatility. If a real-world portfolio could be designed that reliably achieved this profile, presumably many investors would invest in it. The second and third profiles are included mainly to illustrate that *any* function whatsoever may be specified for the profile.

As mentioned above, the portfolios produced by the scheme of Appendix 1 typically include negative weights, which correspond to shorted stocks. As a second set of illustrations, we also analyze portfolios based on the constraint that all weights must be greater than or equal to zero, using the method described in Appendix 2.

The program evaluates the resulting performance of the portfolio's time series by calculating the root-mean-square deviation (RMS) of the portfolio performance from the target profile, namely

$$\mathrm{RMS} = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} \left(\frac{x_i - t_i}{t_i} \right)^2},$$

where x_i is the portfolio time series and t_i is the target profile time series. It also calculates the Sharpe ratio (SR) of the portfolio, in particular

$$\mathrm{SR} = \frac{\sum_{i=1}^{n} (x_i/x_{i-1} - s_i/s_{i-1})}{\sqrt{n \sum_{i=1}^{n} (x_i/x_{i-1} - s_i/s_{k-1})^2}},$$

where s_n is the S&P 500 index time series with reinvested dividends.

Our program implementing the optimization strategy is written in Fortran, using 64-bit

IEEE arithmetic, and incorporates the subroutines `dgefa` and `dgesl` (together with certain lower-level routines) from the Linpack library (Dongarra *et al.*, 2001) for solving the linear system (Equation (A.4)). Running one complete case of results takes only 20 seconds on a MacBook Pro.

5 Results

For our main set of results, we took an in-sample test period $L_1 = 15$ years starting on 31 December 1990 and an out-of-sample test period $L_2 = 10$ years, ending on 4 January 2016. In other words, the in-sample period was 1991 through 2005, and the out-of-sample period was 2006 through 2015. Our program found 277 valid stocks from the S&P 500 database for which data spanning this time period was available. All stock data used here include reinvested dividends.

We show in Figures 1 through 6 plots of the performance of the portfolio, in blue, compared with the target profile in orange. The corresponding profile of the S&P 500 index (with reinvested dividends) is shown in green. The plot on the left of each pair is the standard portfolio, constructed as described in Appendix 1. The plot on the right is for the all-positive-weight portfolio, constructed as described in Appendix 2.

Figures 7 and 8 are the corresponding plots for Profile 2, the stair-step profile. Figures 9 and 10 are the corresponding plots for Profile 3, the sinusoidal profile. These results, with calculated statistics, are summarized in Table 1.

So what are we to make of these results?

Note that in *every* case, the standard portfolio performance achieved zero deviation (to three significant digits) from the target profile over the entire

Table 1 Performance of portfolio generation program versus various target profiles, with in-sample period from 1991 through 2005 (15 years) and out-of-sample period from 2006 through 2015 (10 years).

			Standard weights				All-positive weights			
			RMS dev.		Sharps 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
Sinusoidal	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

"Fig." refers to the figure number; "RMS dev." means root-mean-square deviation from target profile; "Sharpe ratio" means Sharpe ratio relative to S&P 500 with reinvested dividends; "IS" means in-sample; "OOS" means out-of-sample; and "FAIL" means a catastrophic loss of all capital.

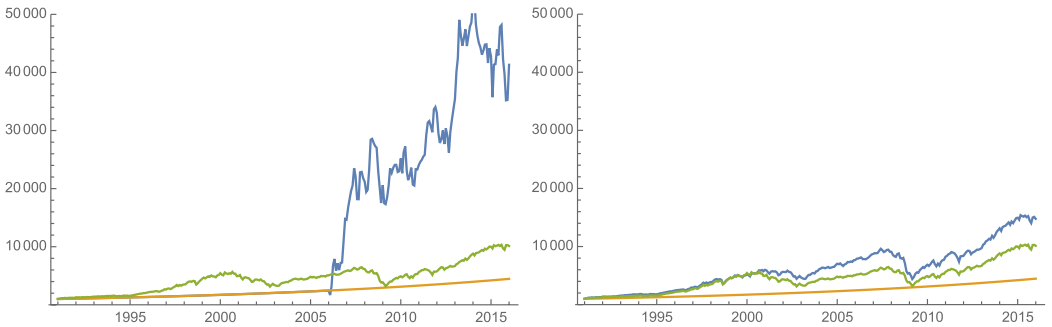

Figure 1 Profile 1, APR = 6%, standard portfolio (L) and all-positive portfolio (R).

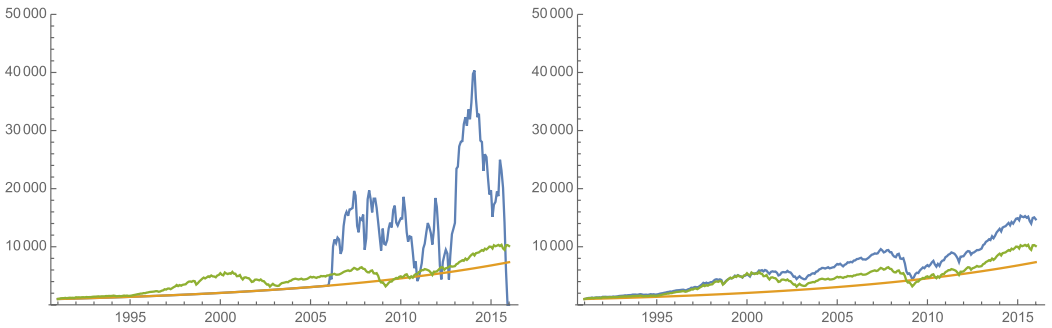

Figure 2 Profile 1, APR = 8%, standard portfolio (L) and all-positive portfolio (R).

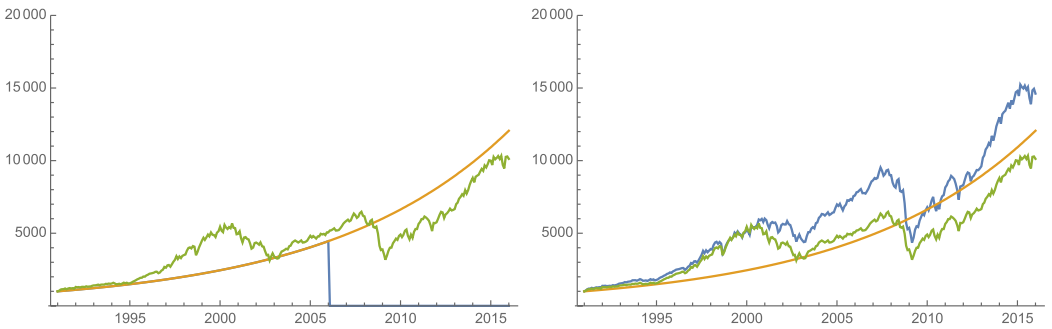

Figure 3 Profile 1, APR = 10%, standard portfolio (L) and all-positive portfolio (R).

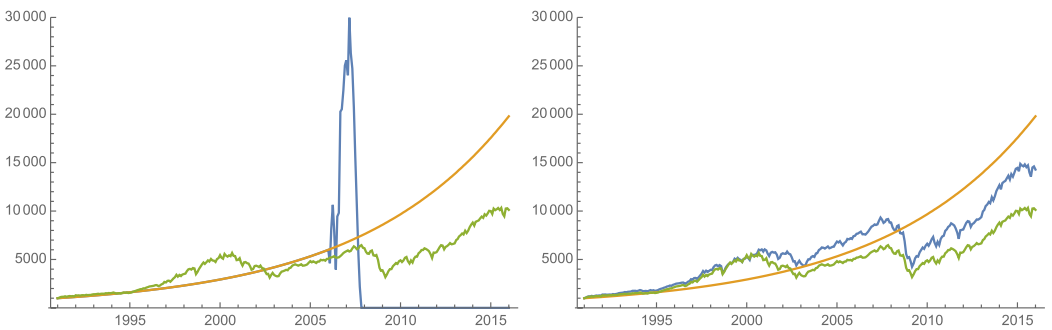

Figure 4 Profile 1, APR = 12%, standard portfolio (L) and all-positive portfolio (R).

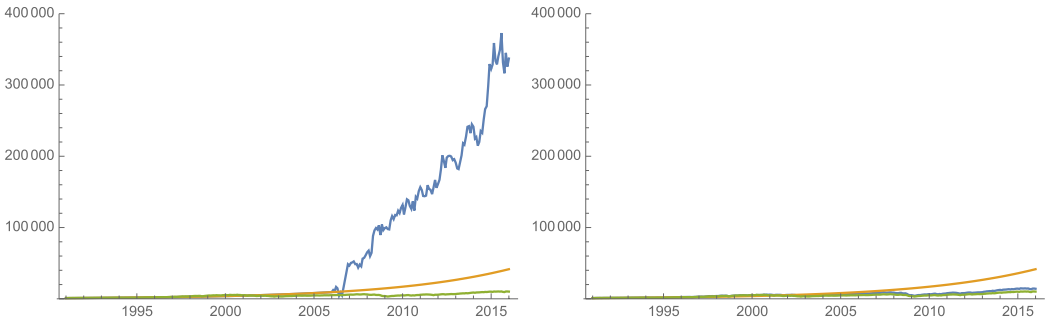

Figure 5 Profile 1, APR = 15%, standard portfolio (L) and all-positive portfolio (R).

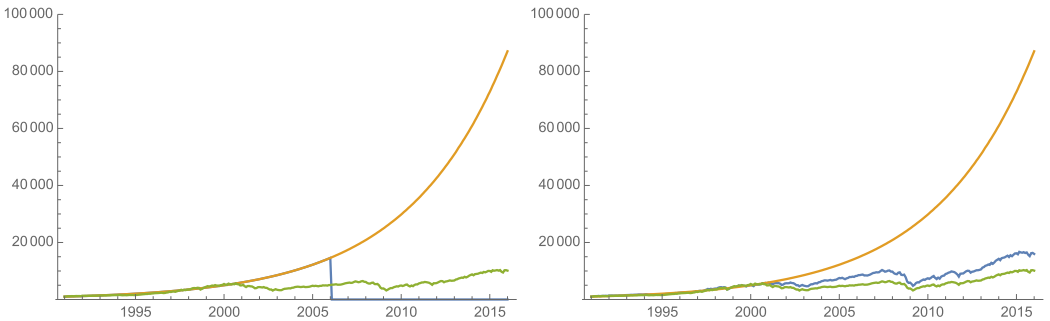

Figure 6 Profile 1, APR = 18%, standard portfolio (L) and all-positive portfolio (R).

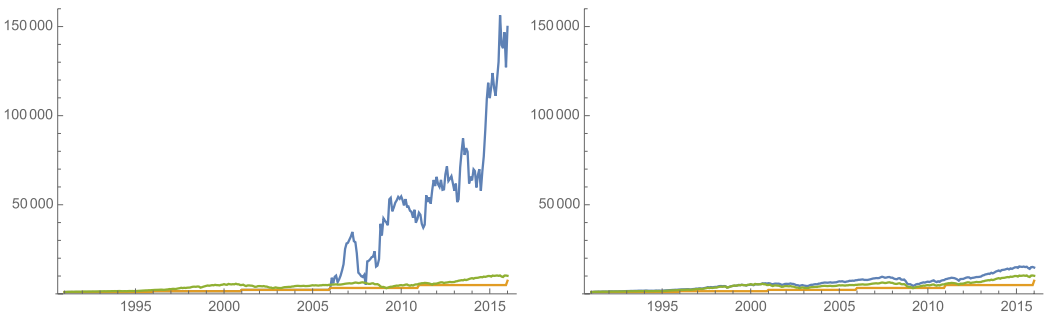

Figure 7 Profile 2, APR = 8%, standard portfolio (L) and all-positive portfolio (R).

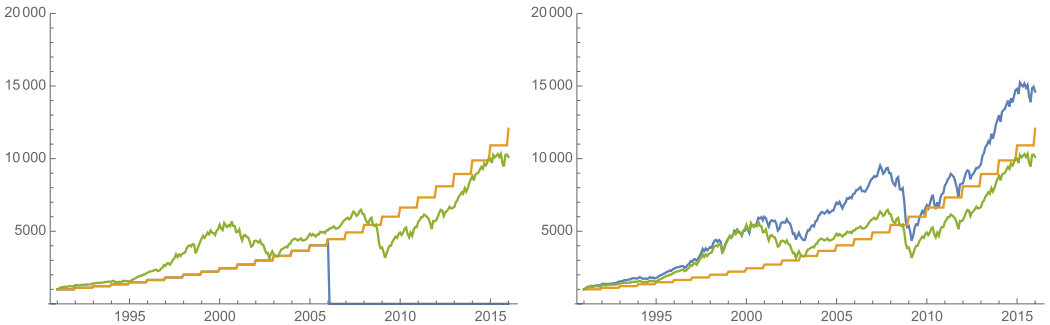

Figure 8 Profile 2, APR = 10%, standard portfolio (L) and all-positive portfolio (R).

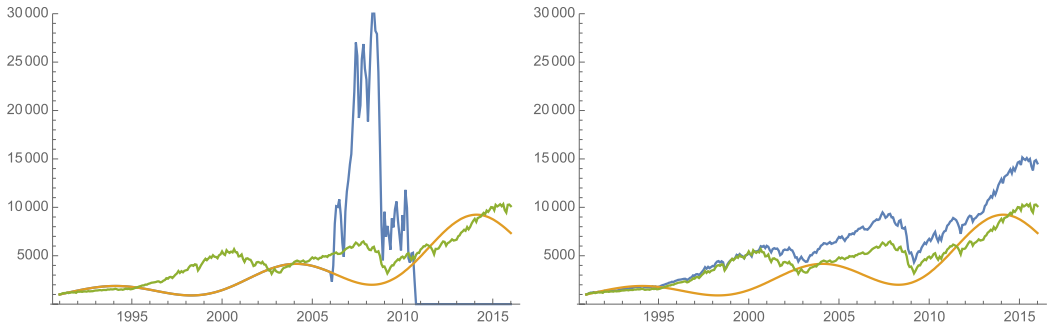

Figure 9 Profile 3, APR = 8%, standard portfolio (L) and all-positive portfolio (R).

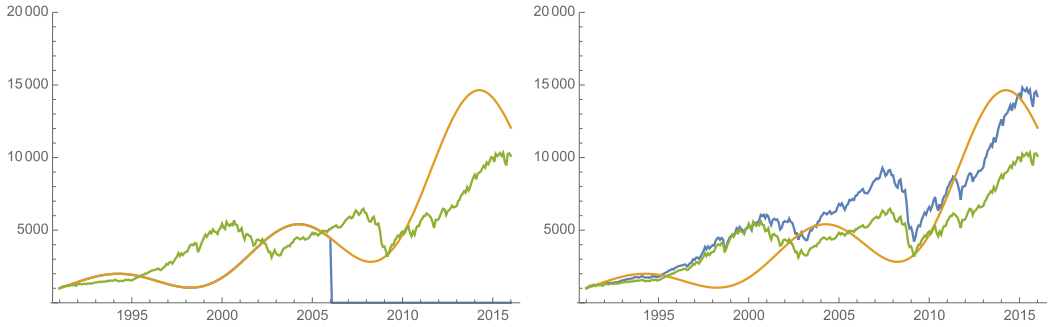

Figure 10 Profile 3, APR = 10%, standard portfolio (L) and all-positive portfolio (R).

15-year in-sample period (1991 through 2005)—an essentially perfect fit in-sample. Only beginning with 2006 (the start of the out-of-sample period) do the blue curves depart from the orange curves in the plots, corresponding to large RMS deviations statistics from the target profile in Table 1.

In some cases, as in Figures 1, 5 and 7, the fitted standard portfolios did remarkably well, far outperforming either the target profile or the S&P 500 index—in some of these cases the vertical axis had to be greatly extended to show the full upward march of the graph (with corresponding Sharpe ratios in Table 1). If one could reliably obtain such performance for future epochs, one could certainly forgive the fact that the performance did not match the target performance profile. But in most other cases, including Figures 2, 3, 4, 6, 8, 9 and 10, the standard portfolios have a very different fate, namely complete ruin—a catastrophic drop to zero, after which the portfolio is presumed to be liquidated (in Table 1 this is indicated by "FAIL" in the out-of-sample Sharpe ratio column).

In any event, it is abundantly clear that the performance of the standard portfolios on out-of-sample data (i.e., beginning with 2006) fails to match the target profiles, as indicated by the much larger RMS deviation statistics. 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, in that they avoid the catastrophic drops to zero that plague the standard portfolios—there are no "FAIL" entries in Table 1 for positive weight cases. In fact, in seven of the ten cases, the positive-weight portfolio outperforms 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. Instead,

these graphs typically resemble scaled (high-beta) versions of the S&P 500 index graphs, with significantly higher volatility than the S&P 500 index.

6 Additional computer runs

We have performed similar analyses for a variety of other cases, with different starting years, in-sample and out-of-sample periods. These are briefly summarized in Table 2. Each row of Table 2 represents a full set of ten runs, corresponding to the ten profile cases as shown in Table 1. A run was deemed a "success" if the out-of-sample test did not suffer a catastrophic loss of all capital and furthermore had a positive Sharpe ratio; otherwise it was deemed a "failure." Sharpe ratios, as before, are based on the S&P 500 index with reinvested dividends.

Table 2 does not present RMS deviation statistics, but none of them were very good—indeed, in the additional runs the RMS deviation statistics were as erratic as in Table 1. The best out-of-sample RMS deviation statistic that we observed in any of these runs was 0.232, for the run starting at 1996 with ten-year IS and OOS periods, with APR = 10%. Most others were considerably higher, indicating only a very weak correlation to the target profile out-of-sample.

With regards to overall performance, some of these additional runs did relatively well. For example, for the case starting in 1986 with 10-year IS and OOS periods, the all-positive-weight portfolio beat the S&P 500 for each of the ten test profiles. The case starting in 1986 with a 15-year IS period and a 10-year OOS period did similarly well with the standard weights. But note that none of these runs achieved uniformly good performance for both the standard and all-positive weight cases.

Also, the runs starting in 1981 or 1986 tend to perform better than those starting in 1991. Keep in mind that data running back before 1991 is now fully 25 years old, and of questionable relevance in today's highly computerized market. Those runs focusing on more recent data did not perform particularly well, suggesting that relatively unsophisticated computer-based portfolio selection schemes such as the ones analyzed in

Table 2 Performance of runs with different starting years and IS periods, in each case with the ten profiles as shown in the rows of Table 1. A run was a "success" if the out-of-sample test did not suffer a catastrophic loss of all capital and had a positive Sharpe ratio; otherwise it was deemed a "failure."

Start	IS	OOS	Standard weights		All-positive weights	
year	years	years	Successes	Failures	Successes	Failures
1981	10	10	1	9	10	0
1986	10	10	0	10	10	0
1991	10	10	1	9	10	0
1996	10	10	0	10	4	6
1981	15	10	8	2	10	0
1986	15	10	9	1	2	8
1991	15	10	3	7	0	10

this study are not likely to perform well in today's market.

7 Overfitting

So what is the best overall explanation for these results? Why do the optimal portfolio performance plots fail so miserably to match those of the target profiles out-of-sample? Here it is worth keeping in mind that the procedure to produce either the standard or all-positive portfolio is tantamount to doing a massive computer search, over the space of all possible sets of weighting factors, for a set of weighting factors that minimizes the R function, either Equation (A.1) or (A.5) respectively.

But consider the size of the space of sets of weighting factors (presuming a positive-weight portfolio): If one allows weights from 0 to 100 percent, in resolution of 0.1 percent, and if there are 100 stocks in the portfolio, then the total number of sets of weighting factors is roughly 100^{1000}, i.e., 10^{2000}, a number vastly greater than the number (10^{86}) of elementary particles in the visible universe. So when a computer program, such as ours, produces an optimal set of weights, it is selecting from an inconceivably large set of possible weighting sets, and thus statistical overfitting of the backtest data is unavoidable.

8 Conclusion

We have shown that it is relatively straightforward to produce a stock portfolio that achieves *any* desired performance profile, based on backtest (in-sample) data. However, the resulting portfolios tend to perform erratically on new (out-of-sample) data, certainly not following the target profile, and, in fact, resulting in complete ruin in many cases. Significantly less erratic results can be obtained by imposing constraints that restrict the portfolio to positive weights, but the resulting portfolios typically depart significantly from the target profile on both the in-sample and out-of-sample data.

The erratic performance observed in our results on out-of-sample data is a classic symptom of statistical overfitting. In fact, overfitting 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.

One way of interpreting our analysis is that selection bias can easily make backtest results worthless. The reason is that as the number of trials grows, so does the probability of selecting a false positive. Readers need to keep in mind that any Sharpe ratio is attainable, given enough trials. Thus there is no such notion as a "minimum" Sharpe ratio that makes a strategy useful. For a detailed discussion of backtest overfitting, we refer the reader to Bailey *et al.* (2014).

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 by trying to discern patterns in stock market indexes by examination of charts (as is often done by technical analysts) or designing a portfolio that tracks a particular risk profile, as many smart beta ETFs attempt. 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, using much more detailed data (minute-by-minute or even millisecond-by-millisecond records of many thousands of securities), who can afford the expertise and facilities to make such analyses profitable. Any lesser efforts, such as those described in this paper, are doomed to be statistically overfit, and if followed may well have disastrous consequences.

Appendix 1: Constructing a weighted portfolio to achieve a desired performance profile

Given a target time series $(v_j, 1 \le j \le n)$ covering a time period T, and a collection of m stocks $(z_i, 1 \le i \le m)$, each with a time series $(z_i(t_j), 1 \le j \le n)$ covering the same time period T, we wish to find m weights $(w_i, 1 \le i \le m)$ for the m stocks that minimize the objective function

$$R(w_1, w_2, \ldots, w_m) = \sum_{j=1}^{n} \left(\sum_{i=1}^{m} w_i z_i(t_j) - v_j \right)^2. \quad \text{(A.1)}$$

Note that $P(t_j) = \sum_{i=1}^{m} w_i z_i(t_j)$ the weighted portfolio time series, so Equation (A.1) is simply the total squared deviation of the portfolio time series from the target time series. The function R is minimized when the following are satisfied:

$$\begin{aligned}
\frac{\partial R}{\partial w_1} &= 2 \sum_{j=1}^{n} \left(\sum_{i=1}^{m} w_i z_i(t_j) - v_j \right) z_1(t_j) = 0, \\
\frac{\partial R}{\partial w_2} &= 2 \sum_{j=1}^{n} \left(\sum_{i=1}^{m} w_i z_i(t_j) - v_j \right) z_2(t_j) = 0, \\
&\cdots \\
\frac{\partial R}{\partial w_m} &= 2 \sum_{j=1}^{n} \left(\sum_{i=1}^{m} w_i z_i(t_j) - v_j \right) z_m(t_j) = 0,
\end{aligned} \quad \text{(A.2)}$$

which can be rewritten as

$$\begin{aligned}
\sum_{i=1}^{m} w_i \sum_{j=1}^{n} z_i(t_j) z_1(t_j) &= \sum_{j=1}^{n} v_j z_1(t_j), \\
\sum_{i=1}^{m} w_i \sum_{j=1}^{n} z_i(t_j) z_2(t_j) &= \sum_{j=1}^{n} v_j z_2(t_j), \\
&\cdots \\
\sum_{i=1}^{m} w_i \sum_{j=1}^{n} z_i(t_j) z_m(t_j) &= \sum_{j=1}^{n} v_j z_m(t_j).
\end{aligned} \quad \text{(A.3)}$$

In matrix notation, Equation (A.3) is the same as writing $A \cdot W = B$, where

$$A = \begin{bmatrix}
\sum_{j=1}^{n} z_1^2(t_j) & \sum_{j=1}^{n} z_1(t_j) z_2(t_j) & \cdots & \sum_{j=1}^{n} z_1(t_j) z_m(t_j) \\
\sum_{j=1}^{n} z_2(t_j) z_1(t_j) & \sum_{j=1}^{n} Z_2^2(t_j) & \cdots & \sum_{j=1}^{n} z_2(t_j) z_m(t_j) \\
\vdots & \vdots & \ddots & \vdots \\
\sum_{j=1}^{n} Z_m(t_j) z_1(t_j) & \sum_{j=1}^{n} Z_m(t_j) z_2(t_j) & \cdots & \sum_{j=1}^{n} Z_2^m(t_j)
\end{bmatrix},$$

$$W = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_m \end{bmatrix}, \quad B = \begin{bmatrix}
\sum_{j=1}^{n} v_j z_1(t_j) \\
\sum_{j=1}^{n} v_j z_2(t_j) \\
\vdots \\
\sum_{j=1}^{n} v_j z_m(t_j)
\end{bmatrix}. \quad \text{(A.4)}$$

In this form, the system can be solved for the W vector by using conventional linear system solver software, such as the Linpack (Dongarra *et al.*, 2001) library, the Lapack (Anderson *et al.*, 1999) library, or even routines built into popular spreadsheet programs. Note that it is not essential that $n > m$; if $n < m$ this scheme produces a best least-squares fit to the target profile, although the quality of this fit degrades when the ratio n/m falls much below one.

Appendix 2: Constructing an all-positive-weight portfolio

As we mentioned above, when the technique described in Appendix 1 is implemented on real stock market data, some of the resulting weights w_i are typically negative. This means that the corresponding stocks will be shorted in the resulting portfolio. While shorting is certainly a legitimate trading strategy, shorting exposes the portfolio to potentially large losses; indeed, this occurred in several examples above. So one can also ask for an

optimal set of weights W subject to the constraint that each weight $w_i > 0$.

Problems of this type are known as optimization problems with interval (or inequality) constraints. These problems have been studied at length, and quite a few sophisticated schemes have been developed for their solution (Bailey and López de Prado, 2013; Barzilai and Borwein, 1998; Bertsekas, 1982; Borwein and Lewis, 1992). In this particular application, we have employed what is known as the logarithmic barrier scheme (Gockenbach, 2003), which is to append a logarithmic term to the minimization problem (Equation (A.1)), as follows:

$$\begin{aligned} &R(w_1, w_2, \ldots, w_m) \\ &\quad = \sum_{j=1}^{n} \left(\sum_{i=1}^{m} w_i z_i(t_j) - v_j \right)^2 \\ &\qquad + 2C \sum_{i=1}^{m} \log w_i. \end{aligned} \tag{A.5}$$

The presence of this logarithmic term penalizes very small weights and thus serves as a barrier, keeping the weights away from zero or negative values. This is not the same as solving the constrained problem, but by successively reducing the constant C, the desired limiting solution can be obtained. In this case, the equivalent minimizing condition is

$$\begin{aligned} \sum_{i=1}^{m} w_i \sum_{j=1}^{n} z_i(t_j) z_1(t_j) &= \sum_{j=1}^{n} v_j z_1(t_j) + C/w_1, \\ \sum_{i=1}^{m} w_i \sum_{j=1}^{n} z_i(t_j) z_2(t_j) &= \sum_{j=1}^{n} v_j z_2(t_j) + C/w_2, \\ &\cdots \\ \sum_{i=1}^{m} w_i \sum_{j=1}^{n} z_i(t_j) z_m(t_j) &= \sum_{j=1}^{n} v_j z_m(t_j) + C/w_m. \end{aligned} \tag{A.6}$$

This system can be efficiently solved by Newton iterations, where one takes, as starting estimates of the weights W, the solution to the unconstrained problem (Equation (A.4)) above, replacing zero or negative weights with some very small positive value. The Newton iterations to be performed are

$$\begin{aligned} \bar{w}_1 &= w_1 - \frac{\sum_{i=1}^{m} w_i \sum_{j=1}^{n} z_i(t_j) z_1(t_j) - \sum_{j=1}^{n} v_j z_1(t_j) - C/w_1}{\sum_{j=1}^{n} z_1^2(t_j) + C/w_1^2}, \\ \bar{w}_2 &= w_2 - \frac{\sum_{i=1}^{m} w_i \sum_{j=1}^{n} z_i(t_j) z_2(t_j) - \sum_{j=1}^{n} v_j z_2(t_j) - C/w_2}{\sum_{j=1}^{n} z_2^2(t_j) + C/w_2^2}, \\ &\cdots \\ \bar{w}_m &= w_m - \frac{\sum_{i=1}^{m} w_i \sum_{j=1}^{n} z_i(t_j) z_m(t_j) - \sum_{j=1}^{n} v_j z_m(t_j) - C/w_m}{\sum_{j=1}^{n} z_m^2(t_j) + C/w_m^2}, \end{aligned} \tag{A.7}$$

In summary, the algorithm for the constrained problem is the following: (a) perform the unconstrained matrix calculation as described in Appendix 1, namely formula (A.4), to obtain an initial set of weights W; (b) replace zero or negative weights with a small positive value (we use 10^{-8}); (c) select $C = 1$, then perform the Newton iteration Equation (A.7) until convergence (typically in ten or fewer iterations); and (d) reduce C by a factor of ten and repeat step (c), continuing until overall convergence (typically when $C = 10^{-6}$ or so).

References

Anderson, E., Bai, Z., Bischof, C., Blackford, S., Demmel, J., Dongarra, J., Du Croz, J., Greenbaum, A., Hammarling, S., McKenney, A., and Sorensen, D. (1999). *LAPACK Users' Guide*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Bailey, D. H. and López de Prado, M. L. (2013). "An Open-Source Implementation of the Critical-Line Algorithm for Portfolio Optimization," *Algorithms* **6**, 169–196.

Bailey, D. H., Borwein, J. M., López de Prado, M. L., 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, pp. 458–471.

Bailey, D. H., Borwein, J. M., Salehipour, A., López de Prado, M. L., and Zhu, Q. J. (2015). "Backtest Overfitting Demonstration Tool: An Online Interface," manuscript, 21 Apr 2015, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2597421.

Barzilai, J. and Borwein, J. M. (1998). "Two-Point Step Size Gradient Methods," *IMA Journal of Numerical Analysis* **8**, 141–148.

Bertsekas, D. P. (1982). "Projected Newton Methods for Optimization Problems with Simple Constrains," *SIAM Journal of Control and Optimization* **20**, 221–246.

Borwein, J. M. and Lewis, A. S. (1992). "Partially Finite Convex Programming, Part I: Quasi Relative Interiors and Duality Theory," *Mathematical Programming* **57**, 15–48.

Chong, E. K. P. and Zak, S. H. (2013). *An Introduction to Optimization*, fourth edition. New York: Wiley.

Dongarra, J. J., Luszczek, P., and Petitet, A. (2001). "The Linpack Benchmark: Past, Present and Future," NETLIB, http://www.netlib.org/utk/people/JackDongarra/PAPERS/hpl.pdf.

Gockenbach, M. S. (2003). "Introduction to Inequality-Constrained Optimization: The Logarithmic Barrier Method," http://www.math.mtu.edu/~msgocken/ma5630spring2003/lectures/bar/bar.pdf.

López de Prado, M. (2016). "Building portfolios that outperform out-of-sample," *Journal of Portfolio Management* **42**(4), 59–69.

Sharpe, W. F. (1994). "The Sharpe Ratio," *Journal of Portfolio Management* **21**, 49–58.

Vanguard Group (Oct 2015). "Exchange-Traded Funds: Clarity Amid the Clutter," https://personal.vanguard.com/pdf/ISGETFC.pdf.

Vanguard Group (Jul 2012). "Joined at the Hip: ETF and Index Development," https://pressroom.vanguard.com/content/nonindexed/7.23.2012_Joined_at_the_hip.pdf.

Keywords: Backtest; historical simulation; backtest overfitting; investment strategy; optimization; Sharpe ratio; minimum backtest length