IPR JOURNALS

inaugural issue

WINTER 2019

VOLUME 1 NUMBER 1

the journal of financial data science

JFDS.iprjournals.com

A Data Science Solution to the Multiple-Testing Crisis in Financial Research

MARCOS LÓPEZ DE PRADO

AQR

A Data Science Solution to the Multiple-Testing Crisis in Financial Research

Marcos López de Prado

Marcos López de Prado
is principal and head of machine learning at AQR Capital Management in Greenwich, CT, and an adjunct professor at Cornell University in Ithaca, NY.
marcos.lopezdeprado@aqr.com

Academics and investors often compute the performance of an investment strategy or factor to determine whether such strategy or factor profits beyond what could be considered "luck." By far the most commonly used investment performance statistic is the Sharpe ratio (SR), first introduced by Sharpe (1966) and further studied by Sharpe (1975, 1994). The probability distribution of this statistic is well known under a variety of assumptions (Lo 2002; Bailey and López de Prado 2012). Using those distributions, it is possible to derive the probability that the observed SR exceeds a given threshold. Under this framework, an investment strategy with a low SR based on a long backtest or track record may be preferred to an alternative strategy with a high SR computed on a short backtest or track record. One problem with this approach is that it does not account for *selection bias under multiple testing* (SBuMT).

In 1933, Jerzy Neyman and Egon Pearson developed the standard hypothesis test used in most scientific applications. These authors did not consider the possibility of performing multiple tests on the same dataset and selecting the most favorable outcome (the one that rejects the null with the lowest false positive probability). At that time, the absence of powerful computers made SBuMT unlikely. Bonferroni (1935) was among the first to recognize that the probability of obtaining a false positive would increase as a test is repeated multiple times over the same dataset. Ever since, statisticians have taken the problem of multiple testing seriously (Gelman and Locken 2013). In its ethical guidelines,[1] the American Statistical Association warns that "failure to disclose the full extent of tests and their results in such a case would be highly misleading" (American Statistical Association 1999).

Given this background, it is surprising to find that practically all papers in empirical finance fail to disclose the number of trials involved in a discovery. Virtually every paper reports a result as if it were the only trial attempted. This is, of course, rarely the case, and it is common for economists to conduct millions of regressions or simulations before finding a result striking enough to merit publication (Sala-i-Martin 1997; Leinweber 2007). Researchers in other fields have taken steps to control for and prevent SBuMT (e.g., visit www.alltrials.net, or see Szucs and Ioannidis 2017). Unlike physics, finance does not have laboratories in which false claims can be easily debunked based on independent tests: All we count on are the same time series used to overfit the backtest, and gathering out-of-sample evidence will take decades (López de Prado 2017).

[1] See Ethical Guideline A.8: http://community.amstat.org/ethics/aboutus/new-item.

A very common misconception is that the problem of SBuMT only affects historical simulations (backtesting). In fact, this problem encompasses any situation in which we select one outcome without controlling for the totality of alternative outcomes from which we choose. For example, a hedge fund may want to hire a portfolio manager with an SR of 2. To that purpose, the fund may interview multiple candidates, not realizing that they should adjust the SR higher with every additional interview. The fact that the SR is computed on an actual track record does not mean that SBuMT will not take place. We could interview a series of dart-throwing monkeys, and eventually we would find one with an SR of 2.

There is nothing wrong with carrying out multiple tests. Researchers should perform multiple tests and report the results of all trials; however, when the extent of the tests carried out is hidden from journal referees, readers, and investors, it is impossible for them to assess whether a particular result is a false positive (Bailey et al. 2014, 2017). For this reason, Harvey, Liu, and Zhu (2016) concluded that "most claimed research findings in financial economics are likely false."

Yet, there is hope. SBuMT can be prevented and corrected in financial economics. Nothing forbids financial researchers from joining the ranks of researchers from other fields who control for SBuMT. Accordingly, the main goal and contribution of this article is to provide a template for how the results from multiple trials could be reported in financial publications. The information regarding all trials could be disclosed in a separate section or an appendix to a publication, while the focus remains on explaining the selected finding. Ideally, the author would report the performance of a proposed investment strategy or factor adjusted for SBuMT. In this particular article we apply the deflated SR (DSR) method (Bailey and López de Prado 2014; López de Prado and Lewis 2018) to control for the effects of SBuMT, non-normality, and sample length. It is not the goal of this article to present a financial discovery or promote an investment strategy, even though the results presented in this publication correspond to an actual investment mandate.

In the following sections, we provide a template for how authors and journals could expose to referees and readers critical information concerning all trials involved in a discovery.

EXHIBIT 1

Performance Statistics for the Index and the Selected Strategy

Statistic	iBoxxIG	Strategy
Start date	1/21/2010	1/21/2010
End date	5/1/2018	5/1/2018
aRoR (Total)	4.90%	9.35%
Avg AUM (1E6)	1,000.00	1,506.43
Avg Gini	0.29	0.88
Avg Duration	7.88	0.08
Avg Default Prob	1.36%	1.58%
An. Sharpe ratio	0.99	2.00
Turnover	0.64	5.68
Effective Number	1034.87	186.26
Correl. to Ix	1.00	0.48
Drawdown (95%)	3.17%	2.89%
Time Underwater (95%)	0.23	0.20
Leverage	1.00	3.59

DISCLOSURE OF ALL TRIALS

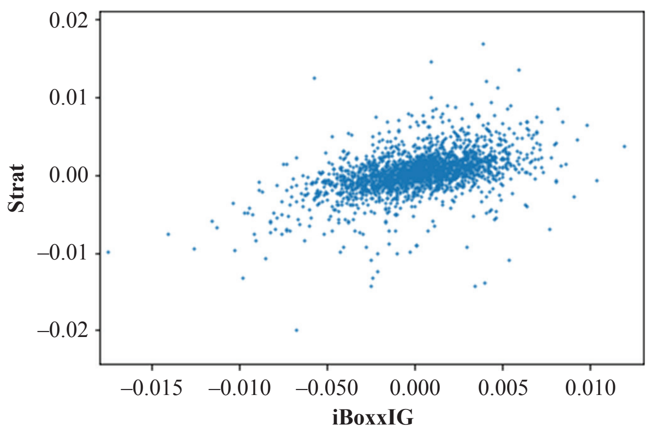
We have developed a market-neutral strategy that invests in liquid high-grade corporate bonds denominated in US dollars. The investment universe is taken from the history of constituents of the Markit iBoxx IG USD index. At each point in time, the strategy may invest in bonds included in the coetaneous index definition, so as to prevent survivorship bias and other forms of information leakage. Although the target portfolio aims at being market neutral, market frictions may prevent all intended trades from being executed. When that happens, the residual risk is hedged with bond futures.

Exhibit 1 lists some statistics associated with the selected strategy. As a reference, it also provides the same information for the index, although results from a long-only index are not directly comparable to those of a market-neutral strategy. Exhibit 2 shows a scatter plot of index returns against strategy returns. The Appendix provides a definition for each of these statistics.

Performance incorporates transaction costs and slippage, based on real transaction cost information collected for this universe over the years. An SR of 2.0 is generally considered high, because the probability of observing that SR after a single trial is infinitesimal, under the null hypothesis that the true SR is zero (see Bailey and López de Prado 2012 for the estimation of such probability).

Exhibit 2

Scatter Plot of iBoxx IG Returns (x-axis) against Strategy Returns (y-axis)



Other specifics about the strategy, such as the underlying principle exploited or predictive features, belong to a different discussion. As explained earlier, our key concern is to provide a template for reporting the information from all trials conducted so that journal referees and investors may evaluate the probability that the discovered strategy is a false positive as a result of SBuMT.

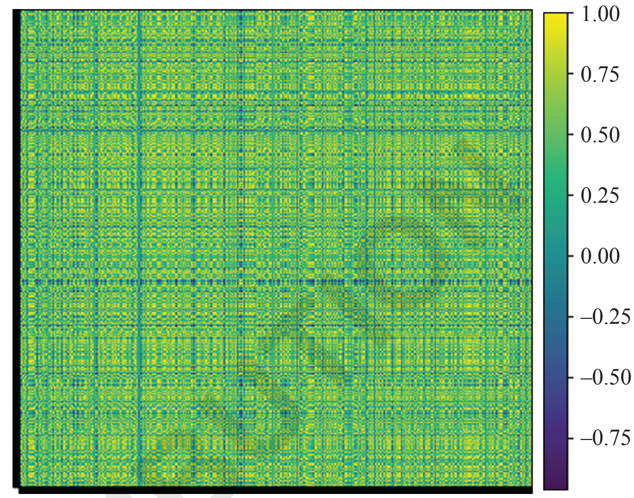
Unlike the practical totality of publications in finance, we begin by acknowledging that the results presented in Exhibits 1 and 2 are not the outcome of a single trial. Because more than one trial took place, the reader must assume that this result is the best out of many alternative ones, and therefore SBuMT is present. By disclosing the information associated with those alternative outcomes, we allow referees and investors to adjust for the inflationary effect of SBuMT.

Exhibit 3 plots the heatmap of return correlations among the 6,385 trials that have taken place before the selection of this investment strategy. This set of trials satisfies the following properties:

- **Complete**
  - The set includes every backtest computed by any of the authors for this or similar investment mandates.
  - Researchers do not have the ability to delete trials, and they are not allowed to backtest outside the official research platform.

Exhibit 3

Heatmap of the Correlation Matrix of the Returns of All 6,385 Trials



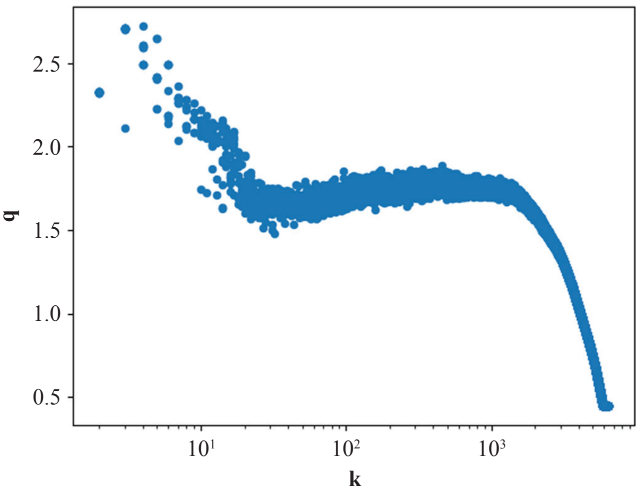
- **Coerced**
  - Researchers do not choose what to log or present. Terabytes of intermediate research metadata are automatically recorded and curated by research surveillance systems.
- **Untainted**
  - Every batch of backtests must be preapproved by the research committee to prevent that externally preselected trials contaminate the internal trials.

External trials are those that have been executed by other authors, outside the control of our research framework. They may have been preselected; hence, they are likely to be biased. To reduce the likelihood of external trials, ideally the research committee may require that trials be justified by *a priori* economic or mathematical theories (e.g., arbitrage-free pricing equations) rather than *a posteriori* empirical theories (e.g., conjectures based on empirical studies).

As is customary in machine learning applications, the main diagonal crosses the Cartesian product from the bottom left to the top right. A light color indicates that the correlation between the returns of two trials was high. The predominance of light colors suggests that the number of uncorrelated trials may be relatively low.

EXHIBIT 4

Quality of Clusters (y-axis) for a Varying Number of Clusters (x-axis, in logarithmic scale)



To assess whether the strategy reported in Exhibit 1 is a false investment strategy, we need to discount the inflationary effect caused by all the trials displayed in Exhibit 3. The first step is to determine the number of effectively uncorrelated clusters of trials.

CLUSTERING OF TRIALS

In this section, we apply the *optimal number of clusters* (ONC) algorithm introduced by López de Prado and Lewis (2018) to the correlation matrix plotted in Exhibit 3. Exhibit 4 plots the measure of the quality of clusters q_k that results from producing k clusters, where $k = 2, \ldots, 6385$. The quality of the clusters seems to collapse beyond $k = 1{,}000$. The higher quality levels are observed for $k < 10$, with the maximum reached by $k = 4$.

Exhibit 5 shows the clustered correlation matrices derived for $k \leq 10$. A visual inspection of these heatmaps seems to confirm that the best clustering is achieved by $k = 4$. For instance, the heatmaps for $k \geq 5$ show multiple large, off-diagonal blocks of highly correlated trials. These off-diagonal blocks appear when very similar trials belong to different (and nonconsecutive) clusters, indicating that the correlation matrix has been overclustered. In contrast, no such off-diagonal blocks can be appreciated in the heatmap for $k = 4$.

One explanation for the low number of clusters is that the researchers tried only strategy configurations that had a rigorous theoretical foundation, derived from mathematical bond pricing equations. The search region was narrowly constrained by predefined mathematical theories. The number of clusters would have been much larger, perhaps in the hundreds, if researchers had tried less mathematical (more arbitrary) configurations.

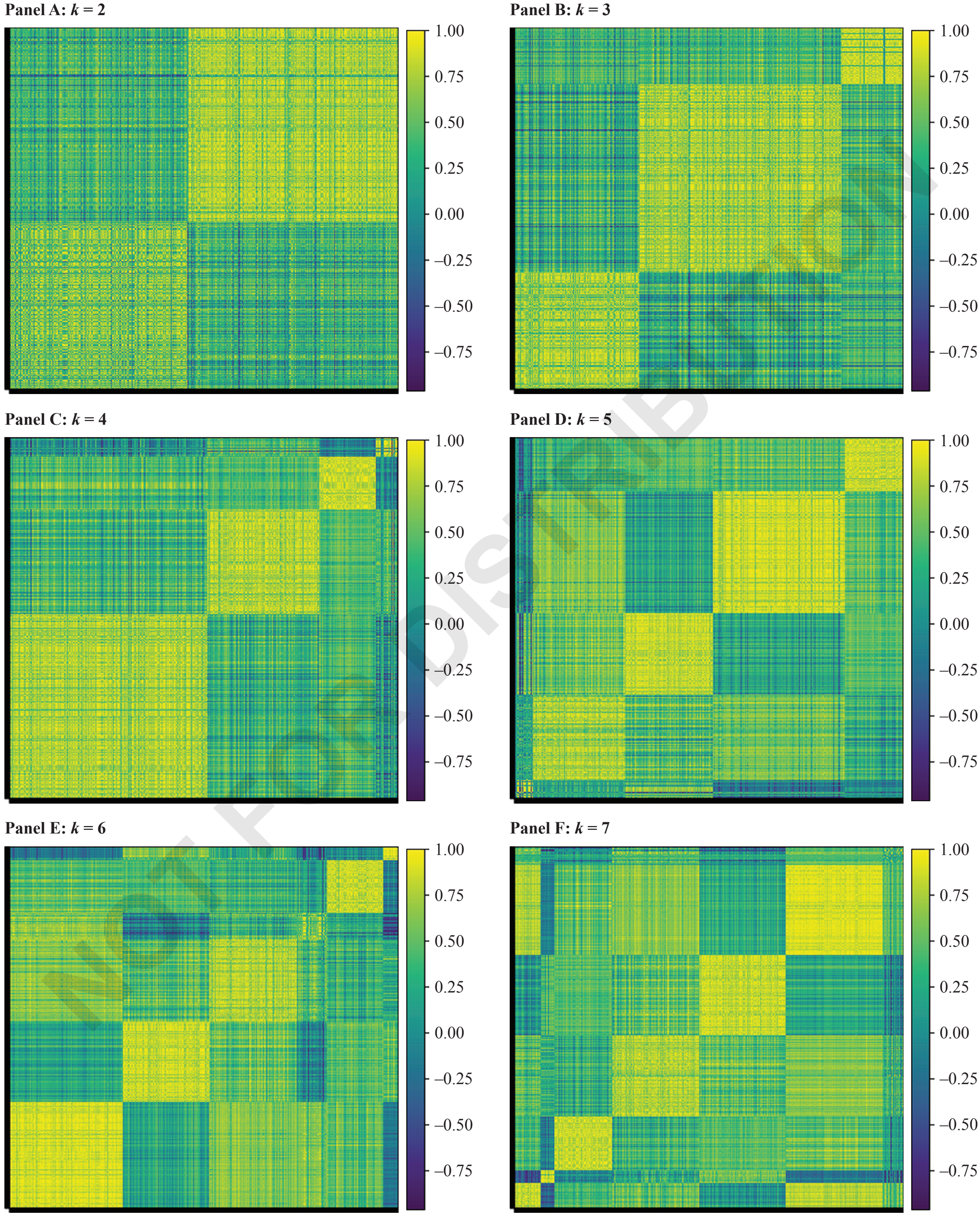
CLUSTER STATISTICS

Following López de Prado and Lewis (2018), we have computed one return series for each cluster; each cluster's composition was determined in the previous section. Forming one time series per cluster further reduces the bias caused by selecting outliers: We do not evaluate the strategy based on a single (potentially "lucky") trial, but based on a large collection of similar trials. In particular, we compute each cluster's returns applying the minimum variance allocation so that highly volatile trials do not dominate the returns time series. Otherwise, a single volatile trial might bias the time series of returns that characterize the entire cluster. Exhibit 6 reports the statistics computed on the clusters' returns series.

For each cluster, we report the following information: (1) *Strat Count* is the number of trials included in a cluster; (2) *aSR* is the annualized SR; (3) *SR* is the nonannualized SR (computed on the same sampling frequency of the original observations; in this case, daily); (4) *Skew* is the skewness of the returns (in the original frequency); (5) *Kurt* is the kurtosis of the returns (in the original frequency); (6) *T* is the number of observations in the returns series; (7) *StartDt* is the date of the first observation in the returns series; (8) *EndDt* is the date of the last observation in the returns series; (9) *Freq* is the average number of observations per year, used to annualize the SR; (10) *sqrt*(*V*[*SR_k*]) is the standard deviation of the SRs across clusters, expressed in the frequency of the cluster; (11) *E*[*max SR_k*] is the expected maximum SR, derived from the false strategy theorem; and (12) *DSR* is the deflated SR—that is, the probability that the true SR exceeds zero after controlling for SBuMT. For the cluster that contains the selected strategy, we have highlighted the *SR* and *E*[*max SR_k*] so that the reader can appreciate the inflationary effect caused by

EXHIBIT 5

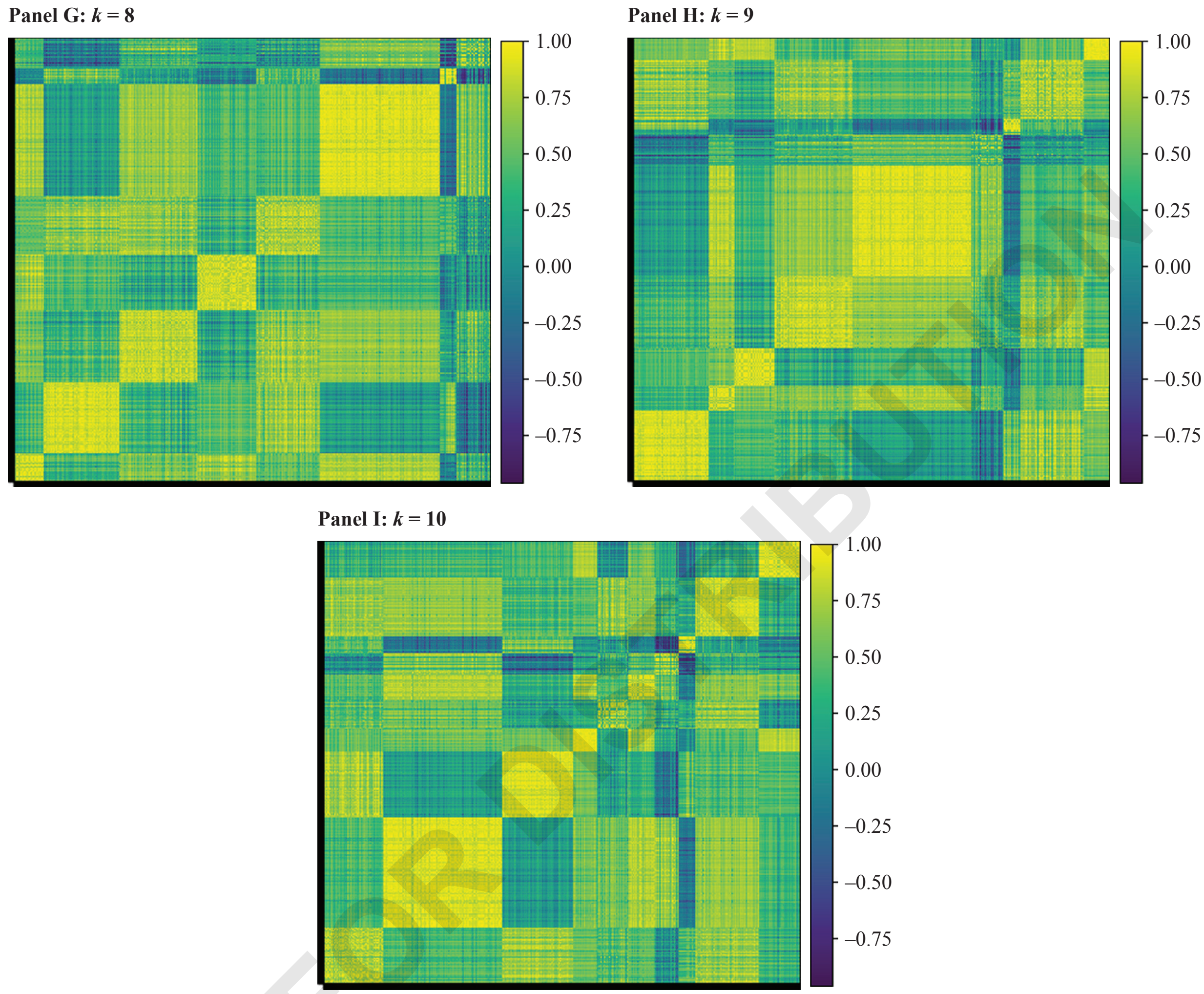
Heatmap of the Clustered Correlation Matrix for $k = 2, \ldots, 10$



(continued)

EXHIBIT 5 *(continued)*

Heatmap of the Clustered Correlation Matrix for k = 2, …, 10



multiple testing. Also highlighted is *DSR*, which corrects for the aforementioned inflation.

Cluster 2 of Exhibit 6 contains the strategy reported in Exhibit 1. The aSR for Cluster 2 is 2.0275, in line with the aSR reported in Exhibit 1. The nonannualized SR is 0.1255, which is consistent with the aSR ($2.0275 \approx 0.1255\sqrt{261.1159}$). Given the number of clusters, and the variance of the cluster SRs, the expected maximum SR (nonannualized) is 0.027, which is significantly lower than 0.1255. Consequently, the DSR is very close to 1. Hence, the probability that the selected strategy is a false positive is virtually zero.

ROBUSTNESS OF THE FINDING

Even though the empirical evidence strongly indicates that $k = 4$ is the optimal clustering, we choose to provide full results for all $k = 2, \ldots, 10$. In this way, referees and readers can evaluate the robustness of the conclusions under alternative scenarios, as unlikely as those scenarios might be. Exhibit 7 displays the cluster statistics for $k = 2,3,5, \ldots, 10$, in the same format we previously used for $k = 4$. For each clustering, we have highlighted the cluster that contains the strategy reported in Exhibit 1.

EXHIBIT 6

Statistics Computed on Clusters' Returns (k = 4, q = 2.7218)

Stats	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Strat Count	3,265	1,843	930	347
aSR	1.5733	1.4907	2.0275	1.0158
SR	0.0974	0.0923	0.1255	0.0629
Skew	–0.3333	–0.4520	–0.4194	0.8058
Kurt	11.2773	6.0953	7.4035	14.2807
T	2,172	2,168	2,174	2,172
StartDt	01-04-2010	01-04-2010	01-04-2010	01-04-2010
EndDt	05-01-2018	04-25-2018	05-03-2018	05-01-2018
Freq	261.0474	261.0821	261.1159	261.0474
sqrt(V[SR_k])	0.0257	0.0256	0.0256	0.0257
E[max SR_k]	0.0270	0.0270	0.0270	0.0270
DSR	0.9993	0.9985	**1.0000**	0.9558

Note: Results for the cluster containing the chosen strategy are shaded.

Results are robust and consistent across all the studied clusterings. The lowest DSR takes place when $k = 10$, where $DSR = 0.9995$. This DSR level is well above the common confidence levels of 0.95 or 0.975 used in most publications. In any event, this DSR corresponds to a very unlikely scenario, given the relatively low quality of the $k = 10$ clustering, compared to the quality achieved by the $k = 4$ clustering. Under these circumstances, we conclude that the strategy underlying these performance results is unlikely to be a false positive caused by SBuMT.

The reader should not infer from this analysis that the strategy will never lose money. All investments involve risk, even those with an SR that almost certainly is positive (see Exhibit 6). The purpose of this analysis was to determine whether the strategy appears to be profitable because of the inflationary effects of SBuMT. Even though the strategy is unlikely to be a false positive, no risky investment can guarantee a positive outcome.

IMPLICATIONS FOR AUTHORS, JOURNALS, AND FINANCIAL FIRMS

The research crisis that afflicts financial economics is not unsolvable. In this article we have presented a template of how this problem can be addressed in practical terms. If the publication of future discoveries could be accompanied with information regarding all the trials involved in those discoveries, financial economics would be able to overcome this crisis and reassert its credibility.

In particular, authors could (1) add to every publication an appendix explaining why the purported discovery is not a false positive caused by SBuMT; (2) certify that they have logged and recorded all the trials that took place during their research; and (3) provide to journal referees the outcomes from all trials. Journals could publish the outcomes from all trials in their websites so that researchers can evaluate the totality of the evidence, not only the trials handpicked by the authors or referees.

Journals could demand that authors (1) disclose all trials; (2) report the extent to which their findings are affected by SBuMT; and (3) evaluate the robustness of their findings to alternative scenarios of SBuMT, as shown in this article.

Financial firms could (1) avoid the practice of optimizing backtests (i.e., picking the winners while ignoring the losers); (2) implement research surveillance frameworks that record, store, and curate every single research trial that takes place within the organization; and (3) estimate the probability of a false positive, objectively controlling for SBuMT.

We believe that adopting these or similar controls for SBuMT would significantly improve the quality of financial journals.

EXHIBIT 7
Statistics Computed on Clusters' Returns

Panel A: $k = 2$, $q = 2.3274$

Stats	Cluster 0	Cluster 1
Strat Count	2,937	3,448
aSR	1.7707	1.6023
SR	0.1096	0.0992
Skew	–0.5780	–0.3351
Kurt	6.5878	11.3212
T	2,174	2,172
StartDt	01-04-2010	01-04-2010
EndDt	05-03-2018	05-01-2018
Freq	261.1159	261.0474
sqrt(V[SR_k])	0.0074	0.0074
E[max SR_k]	0.0038	0.0038
DSR	**1.0000**	1.0000

Panel B: $k = 3$, $q = 2.7068$

Stats	Cluster 0	Cluster 1	Cluster 2
Strat Count	2,063	3,329	993
aSR	1.4411	1.5780	2.0638
SR	0.0892	0.0977	0.1277
Skew	–0.4310	–0.3357	–0.4137
Kurt	5.8606	11.2267	7.3681
T	2,170	2,172	2,174
StartDt	01-04-2010	01-04-2010	01-04-2010
EndDt	04-27-2018	05-01-2018	05-03-2018
Freq	261.1507	261.0474	261.1159
sqrt(V[SR_k])	0.0202	0.0203	0.0202
E[max SR_k]	0.0173	0.0173	0.0173
DSR	0.9995	0.9999	**1.0000**

Panel C: $k = 5$, $q = 2.6517$

Stats	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Strat Count	317	1,524	1,434	2,169	941
aSR	0.9690	1.4664	1.4065	1.5272	2.0319
SR	0.0600	0.0907	0.0870	0.0945	0.1257
Skew	2.2161	–0.3286	–0.4864	–0.4086	–0.4172
Kurt	41.2726	9.7988	5.4162	12.1809	7.4552
T	2,172	2,170	2,168	2,172	2,174
StartDt	01-04-2010	01-04-2010	01-04-2010	01-04-2010	01-04-2010
EndDt	05-01-2018	04-27-2018	04-25-2018	05-01-2018	05-03-2018
Freq	261.0474	261.1507	261.0821	261.0474	261.1159
sqrt(V[SR_k])	0.0234	0.0234	0.0234	0.0234	0.0234
E[max SR_k]	0.0279	0.0279	0.0279	0.0279	0.0279
DSR	0.9418	0.9979	0.9964	0.9987	**1.0000**

(continued)

EXHIBIT 7 *(continued)*
Statistics Computed on Clusters' Returns

Panel D: k = 6, q = 2.4919

Stats	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Strat Count	1,873	1,418	1,447	476	935	236
aSR	1.5205	1.4034	1.4580	1.3853	2.0296	0.4322
SR	0.0941	0.0869	0.0902	0.0857	0.1256	0.0267
Skew	–0.4254	–0.4872	–0.3458	0.5432	–0.4188	0.1344
Kurt	13.0185	5.4077	9.9281	16.1401	7.4308	5.6976
T	2,170	2,168	2,170	2,172	2,174	2,170
StartDt	01-04-2010	01-04-2010	01-04-2010	01-04-2010	01-04-2010	01-04-2010
EndDt	04-27-2018	04-25-2018	04-27-2018	05-01-2018	05-03-2018	04-27-2018
Freq	261.1507	261.0821	261.1507	261.0474	261.1159	261.1507
sqrt(V[SR_k])	0.0321	0.0321	0.0321	0.0321	0.0321	0.0321
E[max SR_k]	0.0417	0.0418	0.0417	0.0418	0.0417	0.0417
DSR	0.9909	0.9797	0.9862	0.9807	**0.9999**	0.2421

Panel E: k = 7, q = 2.3650

Stats	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Strat Count	443	232	940	1,436	1,418	1,591	325
aSR	1.4985	0.4229	2.0314	1.4566	1.4034	1.4816	1.2380
SR	0.0927	0.0262	0.1257	0.0901	0.0869	0.0917	0.0766
Skew	–0.4098	0.1355	–0.4174	–0.3447	–0.4872	–0.4488	10.2898
Kurt	10.4565	5.6820	7.4499	9.9064	5.4077	13.8743	295.3934
T	2,170	2,170	2,174	2,169	2,168	2,170	2,172
StartDt	01-04-2010	01-04-2010	01-04-2010	01-04-2010	01-04-2010	01-04-2010	01-04-2010
EndDt	04-27-2018	04-27-2018	05-03-2018	04-26-2018	04-25-2018	04-27-2018	05-01-2018
Freq	261.1507	261.1507	261.1159	261.1164	261.0821	261.1507	261.0474
sqrt(V[SR_k])	0.0298	0.0298	0.0298	0.0298	0.0298	0.0298	0.0298
E[max SR_k]	0.0413	0.0413	0.0413	0.0413	0.0413	0.0413	0.0413
DSR	0.9901	0.2403	**0.9999**	0.9868	0.9807	0.9884	0.9799

Panel F: k = 8, q = 2.2822

Stats	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Strat Count	411	1,021	1,037	794	846	1,606	228	442
aSR	1.8643	1.3267	1.4133	1.9881	1.5228	1.4607	0.3817	1.3586
SR	0.1154	0.0821	0.0875	0.1230	0.0942	0.0904	0.0236	0.0841
Skew	–0.2217	–0.4884	–0.3657	–0.4156	–0.3822	–0.4481	0.1270	1.6051
Kurt	13.2850	5.1541	10.3922	6.7874	7.4346	12.7538	5.3075	34.8674
T	2,170	2,167	2,169	2	2,168	2,170	2,170	2,172
StartDt	01-04-2010	01-05-2010	01-04-2010	01-04-2010	01-04-2010	01-04-2010	01-04-2010	01-04-2010
EndDt	04-27-2018	04-25-2018	04-26-2018	05-03-2018	04-25-2018	04-27-2018	04-27-2018	05-01-2018
Freq	261.1507	261.0477	261.1164	261.1159	261.0821	261.1507	261.1507	261.0474
sqrt(V[SR_k])	0.0298	0.0298	0.0298	0.0298	0.0298	0.0298	0.0298	0.0298
E[max SR_k]	0.0435	0.0435	0.0435	0.0435	0.0435	0.0435	0.0435	0.0435
DSR	0.9994	0.9606	0.9772	**0.9998**	0.9895	0.9829	0.1774	0.9754

(continued)

EXHIBIT 7 *(continued)*

Statistics Computed on Clusters' Returns

Panel G: $k = 9$, $q = 2.2594$

Stats	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8
Strat Count	1,021	352	536	1,037	1,593	440	228	846	332
aSR	1.3267	1.8185	1.8971	1.4133	1.4578	1.3482	0.3817	1.5228	1.9497
SR	0.0821	0.1125	0.1174	0.0875	0.0902	0.0834	0.0236	0.0942	0.1207
Skew	–0.4884	–0.2077	–0.3769	–0.3657	–0.4467	2.2752	0.1270	–0.3822	–0.4008
Kurt	5.1541	13.3085	6.1852	10.3922	12.7629	49.3210	5.3075	7.4346	10.0715
T	2,167	2,170	2,160	2,169	2,170	2,172	2,170	2,168	2,171
StartDt	01-05-2010	01-04-2010	01-22-2010	01-04-2010	01-04-2010	01-04-2010	01-04-2010	01-04-2010	01-04-2010
EndDt	04-25-2018	04-27-2018	05-03-2018	04-26-2018	04-27-2018	05-01-2018	04-27-2018	04-25-2018	04-30-2018
Freq	261.0477	261.1507	260.9792	261.1164	261.1507	261.0474	261.1507	261.0821	261.0131
sqrt(V[SR_k])	0.0290	0.0290	0.0290	0.0290	0.0290	0.0290	0.0290	0.0290	0.0290
E[max SR_k]	0.0441	0.0441	0.0441	0.0441	0.0441	0.0441	0.0441	0.0441	0.0441
DSR	0.9580	0.9990	**0.9995**	0.9755	0.9813	0.9736	0.1696	0.9886	0.9997

Panel H: $k = 10$, $q = 2.2211$

Stats	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9
Strat Count	806	1,596	948	332	409	353	327	227	851	536
aSR	1.5222	1.4586	1.3083	1.9497	1.3378	1.8174	1.2172	0.3787	1.4057	1.8971
SR	0.0942	0.0903	0.0810	0.1207	0.0828	0.1125	0.0753	0.0234	0.0870	0.1174
Skew	–0.3953	–0.4461	–0.4847	–0.4008	–0.1356	–0.2065	4.5167	0.1274	–0.4064	–0.3769
Kurt	6.9109	12.7512	5.1189	10.0715	7.4999	13.3321	108.1831	5.3035	10.9871	6.1852
T	2,168	2,170	2,167	2,171	2,170	2,170	2,172	2,170	2,169	2,160
StartDt	01-04-2010	01-04-2010	01-05-2010	01-04-2010	01-04-2010	01-04-2010	01-04-2010	01-04-2010	01-04-2010	01-22-2010
EndDt	04-25-2018	04-27-2018	04-25-2018	04-30-2018	04-27-2018	04-27-2018	05-01-2018	04-27-2018	04-26-2018	05-03-2018
Freq	261.0821	261.1507	261.0477	261.0131	261.1507	261.1507	261.0474	261.1507	261.1164	260.9792
sqrt(V[SR_k])	0.0278	0.0278	0.0278	0.0279	0.0278	0.0278	0.0278	0.0278	0.0278	0.0279
E[max SR_k]	0.0438	0.0438	0.0439	0.0439	0.0438	0.0438	0.0439	0.0438	0.0438	0.0439
DSR	0.9889	0.9819	0.9544	0.9997	0.9636	0.9990	0.9483	0.1706	0.9748	**0.9995**

Note: Results for the cluster containing the chosen strategy are shaded.

ACKNOWLEDGMENTS

Special thanks to Prof. Riccardo Rebonato for his valuable comments and to Dr. Michael J. Lewis for his assistance with the exhibits.

APPENDIX

PERFORMANCE STATISTICS

aRoR (Total)

Total return obtained by annualizing the geometrically linked total daily returns. This includes returns due to income from coupons, clean price changes, and financing.

Avg AUM (1E6)

Average of the daily assets under management of the long portfolio, expressed in millions of US dollars.

Avg Gini

Average of the daily Gini coefficients. The daily Gini coefficient is the ratio (1) and (2), where (1) is the area between the Lorenz curve and the line of equality and (2) is the area under the line of equality. The input is the vector of allocations (w) for the ISINs in the index at that moment.

```
def getGiniCoeff(w):
    w=w/w.sum()
    N=len(w)
    Ideal=(N+1)/2.
    lorenz=np.sum(np.cumsum(np.sort(w)))
    return (ideal-lorenz)/ideal
```

Avg Duration

Average of the daily weighted average durations of the portfolio (includes long, short, and futures positions), where the weights are derived from market value allocations. The daily weighted average duration δ_t is computed as

$$\delta_t = \frac{\sum_{k=0}^{n} \omega_{t,n} \delta_{t,n}}{\sum_{k=0}^{n} \left| \omega_{t,n} \right|}$$

Avg Default Prob

Average of the daily weighted average default probabilities of long positions. Weights are derived from market value allocations. A default on a short position is favorable; hence, only long positions are included in the calculation.

An. Sharpe Ratio

Annualized Sharpe ratio computed from daily total returns.

Turnover

Annualized turnover measures the ratio of the average dollar amount traded per year to the average annual assets under management.

Effective Number

The effective number of positions in the portfolio, controlling for concentration of allocations. For a detailed explanation, see López de Prado (2018), Chapter 18, Section 18.7.

```
def getEffNum(w):
    w=w.replace(0,np.nan)
    return np.exp(-(w*np.log(w)).sum())
```

Correl to Ix

Correlation of daily returns relative to the index.

Drawdown (95%)

The 95th percentile across all drawdowns. Drawdowns are computed using the following function.

```
def computeDD_TuW(series,dollars=False):
    df0=series.to_frame('pnl')
    df0['hwm']=series.expanding().max()
    df1=df0.groupby('hwm').min().reset_index()
    df1.columns=['hwm','min']
    df1.index=df0['hwm'].drop_duplicates \
        (keep='first').index # time of hwm
    df1=df1[df1['hwm']>df1['min']]
    if dollars:dd=df1['hwm']−df1['min']
    else:dd=1−df1['min']/df1['hwm']
    tuw=((df1.index[1:]−df1.index[:−1])/ \
        np.timedelta64(1,'Y')).values # in years
    tuw=pd.Series(tuw,index=df1.index[:−1])
    return dd,tuw
```

Time Underwater (95%)

The 95th percentile across all time underwater. The series of time underwater is computed using the above function.

Leverage

Average of the daily leverage. Daily leverage is defined as the ratio between the market value of the long positions and the assets under management.

REFERENCES

American Statistical Association. 1999. "Ethical Guidelines for Statistical Practice." Committee on Professional Ethics, approved by the Board of Directors (August 7, 1999). http://community.amstat.org/ethics/aboutus/new-item.

Bailey, D., J. Borwein, M. López de Prado, and J. Zhu. 2014. "Pseudo-Mathematics and Financial Charlatanism: The Effects of Backtest Overfitting on Out-of-Sample Performance." *Notices of the American Mathematical Society* 61 (5): 458–471.

——. 2017. "The Probability of Backtest Overfitting." *Journal of Computational Finance* 20 (4): 39–70.

Bailey, D., and M. López de Prado. 2012. "The Sharpe Ratio Efficient Frontier." *Journal of Risk* 15 (2): 3–44.

——. 2014. "The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting and Non-Normality." *The Journal of Portfolio Management* 40 (5): 94–107.

Bonferroni, C. E. 1935. "Il calcolo delle assicurazioni su gruppi di teste." Tipografia del Senato.

Gelman, A., and E. Locken. 2013. "The Garden of Forking Paths: Why Multiple Comparisons Can Be a Problem, Even When There Is No 'Fishing Expedition' or 'p-hacking' and the Research Hypothesis Was Posited Ahead of Time." Working paper, http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf.

Harvey, C., Y. Liu, and C. Zhu. 2016. "…and the Cross-Section of Expected Returns." *Review of Financial Studies* 29 (1): 5–68.

Leinweber, D. 2007. "Stupid Data Miner Tricks: Overfitting the S&P 500." *The Journal of Investing* 16 (1): 15–22.

Lo, A. 2002. "The Statistics of Sharpe Ratios." *Financial Analysts Journal* (July): 36–52.

López de Prado, M. 2017. "Finance as an Industrial Science." *The Journal of Portfolio Management* 43 (4): 5–9.

——. *Advances in Financial Machine Learning*, 1st ed. Hoboken, NJ: Wiley, 2018.

López de Prado, M., and M. Lewis. 2018. "Detection of False Investment Strategies Using Unsupervised Learning Methods." Working paper, https://ssrn.com/abstract=3167017.

Sala-i-Martin, X. 1997. "I Just Ran Two Million Regressions." *American Economic Review* 87 (2).

Sharpe, W. 1966. "Mutual Fund Performance." *The Journal of Business* 39 (1): 119–138.

——. 1975. "Adjusting for Risk in Portfolio Performance Measurement." *The Journal of Portfolio Management* 1 (2): 29–34.

——. 1994. "The Sharpe Ratio." *The Journal of Portfolio Management* 21 (1): 49–58.

Szucs, D., and J. Ioannidis. 2017. "When Null Hypothesis Significance Testing Is Unsuitable for Research: A Reassessment." *Frontiers in Human Neuroscience* 11 (Article 390).

Disclaimer

AQR Capital Management is a global investment management firm and may or may not apply investment techniques or methods of analysis similar to those described herein. The views expressed here are those of the authors and not necessarily those of AQR.