A PRACTICAL SOLUTION TO THE MULTIPLE-TESTING CRISIS IN FINANCIAL RESEARCH

Marcos López de Prado [a,b]
Lee Cohn [b]
Michael J. Lewis [b]
Michael Lock [b]
Yaxiong Zeng [b]
Zhibai Zhang [b]

First version: May 4, 2018
This version: May 11, 2018

[a] Research Fellow, Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720. E-mail: lopezdeprado@lbl.gov. Web: www.QuantResearch.org
[b] True Positive Technologies, LP. New York, NY 10017. Web: www.TruePositive.com

A PRACTICAL SOLUTION TO THE MULTIPLE-TESTING CRISIS IN FINANCIAL RESEARCH

ABSTRACT

Most discoveries in empirical finance are false, as a consequence of selection bias under multiple testing. This may explain why so many hedge funds fail to perform as advertised or as expected. These false discoveries may have been prevented if academic journals and investors demanded that any reported investment performance incorporates the false positive probability, adjusted for selection bias under multiple testing. In this paper, we present a real example of how this adjusted false positive probability can be computed and reported for public consumption.[1]

Keywords: Backtest overfitting, selection bias, multiple testing, quantitative investments, machine learning, financial fraud, smart beta, factor investing.

JEL Classification: G0, G1, G2, G15, G24, E44.
AMS Classification: 91G10, 91G60, 91G70, 62C, 60E.

[1] Notice: Special thanks to Prof. Riccardo Rebonato for his valuable comments. Some of the methods discussed in this paper are partly based on the book *Advances in Financial Machine Learning* (Wiley, 2018), available at https://goo.gl/w6gMdq. True Positive Technologies LP has filed multiple patent applications covering business processes and methods that prevent the occurrence of false discoveries in finance, including U.S. Patent Applications No. 14/672,028, No. 15/904,523, No. 62/646,421, No. 62/649,633, and International Application No. PCT/US2015/023198. This publication is intended for academic discussion only. It does not constitute investment advice, and it does not recommend a particular course of action. The opinions expressed here are solely the authors', and they do not necessarily reflect those of the organizations they are affiliated with.

1. INTRODUCTION

Academics and investors often compute the performance of an investment strategy or factor, in order 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 in Sharpe [1966] and further studied in 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. In its ethical guidelines,[2] 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 was 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]). Finance may be the last remaining field oblivious to this methodological error, as researchers in other fields have taken steps to control for and prevent SBuMT (e.g., visit www.alltrials.net, see Szucs and Ioannidis [2017]). One reason why finance has gotten away for so long with this research fraud is that we do not have laboratories where false claims can be debunked based on new evidence: 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]).

A very common misconception is that the problem of SBuMT only affects historical simulations (backtesting). In fact, this problem encompasses any situation where we select one outcome, without controlling for the totality of alternative outcomes we chose from. For example, a hedge fund may want to hire a portfolio manager with a 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 will find one with a SR of 2.

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

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. For this reason, Harvey et al. [2016] concluded that "most claimed research findings in financial economics are likely false." Think about the profound implications of this statement. First, we could burn the entire literature of empirical research in finance, including many papers written by Nobel laureates and tenured professors, and sadly, the loss to the subject may be negligible. We are not aware of a single journal article in the vast "factor investing" literature that has reported or adjusted for all trials. Second, trillions of dollars are invested in funds and financial products based on these false discoveries, like "smart beta" funds. Investors pay tens of billions of dollars in fees every year, even though these investments do not perform as advertised or expected, because customers have been misled to believe that these are scientific products. The reality is, these firms are taking advantage of the public's trust in science, with the tacit approval of the academic community.

The public may soon realize that empirical finance is not a field of scientific research, because blatant disregard for SBuMT has led to the widespread proliferation of false positives. Unless the problem of SBuMT is addressed, empirical finance will be considered a pseudo-science, at par with astrology, alchemy or medical quackery. Astrologers follow precise and complex rules to produce their horoscopes, and so do investment advisers and econometricians, with similar degrees of failure and selection bias. What makes empirical finance a pseudo-science is not that it expresses opinions or beliefs, but that these opinions or beliefs are misrepresented as statements of fact, falsely backed by anecdotal empirical evidence. There are many legitimate academic fields, like philosophy or theology, which do not claim to be scientific. In contrast, empirical finance aspires to be a science without abiding by the rules of science.

Yet, there is hope. SBuMT can be prevented and corrected in financial economics. Nothing forbids financial researchers from joining the ranks of legitimate researchers from other fields who control for SBuMT. Accordingly, the main goal and contribution of this paper 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 exposed in a separate section or an appendix to a publication, while the focus remains on explaining the selected finding. Ideally, the author should report the performance of a proposed investment strategy or factor adjusted for SBuMT. In this particular paper we apply the deflated Sharpe ratio (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 paper to present a financial discovery or promote an investment strategy, even though the results presented in this publication correspond to an actual investment mandate.

The rest of the paper is structured as follows: Section 2 illustrates how authors could disclose the information concerning all trials involved in the discovery of a particular investment strategy. Section 3 lists a number of steps that authors, journals and financial firms could take in order to overcome the current research crisis. Section 4 summarizes our conclusions. The Appendix defines the terms used to characterize the performance of a strategy. A section titled "Frequently Asked Questions" comments on some interesting ideas raised by readers of earlier versions of this manuscript.

2. A TEMPLATE FOR CONTROLLING FOR SBuMT

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

2.1. DISCLOSURE OF ALL TRIALS

We have developed a market neutral strategy that invests in liquid high grade corporate bonds denominated in U.S. 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 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 1.1 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 1.2 shows a scatter plot of index returns against strategy returns. Appendix A.1 provides a definition for each of these statistics.

[EXHIBIT 1 HERE]

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

Other specifics about the strategy, like the underlying principle exploited or data sources, 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 Exhibit 1 are not the outcome of a single trial. Since more than one trial took place, the reader must assume that this result is the best out of many alternative ones, and therefore selection bias is present. By disclosing the information associated with those alternative outcomes, we allow referees and investors to adjust for the inflationary effect of selection bias.

[EXHIBIT 2 HERE]

Exhibit 2 plots the heatmap of returns correlation between 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.
- **Coerced**:
    - Researchers do not choose what to log or present. Terabytes of intermediate research meta-data are automatically recorded and curated by research surveillance systems.
- **Untainted**:
    - Every batch of backtests must be pre-approved by the research committee, in order to prevent that external trials could 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 pre-selected, hence they are likely to be biased. In order to reduce the likelihood of external trials, ideally the research committee may require that trials are justified by *a priori* mathematical theories (such as arbitrage-free pricing equations) rather than *a posteriori* empirical theories (such as conjectures based upon empirical studies).

As it 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.

In order 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 2. The first step is to determine the number of essentially uncorrelated clusters of trials.

2.2. CLUSTERING OF TRIALS

In this section, we apply the base clustering algorithm explained in López de Prado and Lewis [2018] to the correlation matrix plotted in Exhibit 2. Exhibit 3 plots the measure of quality of clusters q_k that result from producing k clusters, where $k = 2, \ldots, 6384$. The quality of the clusters seems to collapse beyond $k = 1000$. The highest quality is observed for $k < 10$, with the maximum reached by $k = 4$.

[EXHIBIT 3 HERE]

Exhibit 4 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 non-consecutive) clusters, indicating that the correlation matrix has been over-clustered. In contrast, no such off-diagonal blocks can be appreciated in the heatmap for $k = 4$.

[EXHIBIT 4 HERE]

One explanation for the low number of clusters is that the researchers only tried 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, like the ones often found in the economic and factor investing literature.

2.3. CLUSTER STATISTICS

Following López de Prado and Lewis [2018], we have computed one return series for each cluster, where each cluster's composition was determined in the previous section. Forming one times series per cluster further reduces the bias caused by selecting outliers, because 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 time series. Otherwise, a single volatile trial might bias the time series of returns that characterize the entire cluster. Exhibit 5 reports the statistics computed on the clusters' returns series.

[EXHIBIT 5 HERE]

For each cluster, we report the following information: (i) *Strat Count* is the number of trials included in a cluster; (ii) *aSR* is the annualized SR; (iii) *SR* is the non-annualized SR (computed on the same sampling frequency of the original observations, in this case daily); (iv) *Skew* is the skewness of the returns (in the original frequency); (v) *Kurt* is the kurtosis of the returns (in the original frequency); (vi) *T* is the number of observations in the returns series; (vii) *StartDt* is the date of the first observation in the returns series; (viii) *EndDt* is the date of the last observation in the returns series; (ix) *Freq* is the average number of observations per year; (x) *sqrt(V[SR_k])* is the standard deviation of the SRs across clusters, expressed in the frequency of the cluster; (xi) *E[max SR_k]* is the expected maximum SR, derived from the "False Strategy" theorem; (xii) *DSR* is the deflated SR, i.e. the probability that the true SR exceeds zero after controlling for SBuMT.

Cluster 2 of Exhibit 5 contains the strategy reported in Exhibit 1. The annualized SR for Cluster 2 is 2.0275, in line with the annualized SR reported in Exhibit 1. The non-annualized SR is 0.1255, which is consistent with the annualized SR ($2.0275 \approx 0.1255\sqrt{261.1159}$). Given the number of clusters, and the variance of the cluster SRs, the expected maximum SR (non-annualized) is 0.027, which is significantly lower than 0.1255. Consequently, the DSR is very close to 1.

2.4. 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 6 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 in yellow the cluster that contains the strategy reported in Exhibit 1.

[EXHIBIT 6 HERE]

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 using 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. In all cases, $DSR > 0.99$. 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 a SR that almost certainly is positive (see Exhibit 5). The purpose of this analysis was to determine whether the strategy appears to be profitable due to the inflationary effects of SBuMT. Even though the strategy is unlikely to be a false positive, no risky investment can guarantee a positive outcome.

3. IMPLICATIONS FOR AUTHORS, JOURNALS AND FINANCIAL FIRMS

The research crisis that afflicts financial economics is not unsolvable. In this paper we have presented a template of how this problem can be solved 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 regain the credibility it has lost.

In particular, authors should: (i) Add to every publication an appendix explaining why the purported discovery is not a false positive caused by SBuMT; (ii) certify that they have logged and recorded all the trials that took place during their research; and (iii) provide to journal referees the outcomes from all trials. Journals must 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.

Financial firms should: (i) Stop the dishonest practice of optimizing backtests, picking the winners while concealing the losers; (ii) cease to commercialize funds and products based on research where authors did not control for all trials; (iii) implement research surveillance frameworks that record, store and curate every single research trial that takes place within the organization; and (iv) estimate the probability of a false positive, controlling for SBuMT, for every new product.

4. CONCLUSIONS

The peer-review process of research in financial economics is broken, for the reasons stated in the introduction to this paper. Our hope with this publication is that, going forward, financial economics will join other fields of research, and take seriously the problem of SBuMT. Nothing less than the credibility of its entire body of work is at stake.

The consequences from this crisis reach far beyond University campuses. A myriad of financial products is based on false discoveries published in financial journals over the past decades. Investors have paid the price for these false discoveries, which can be quantified in terms of loss of principal investments, but also in terms of unjustifiable fees for no service, and the opportunity cost of misallocating assets.

Investors should stop purchasing financial products based on false discoveries, where academic journals have not controlled for selection bias. If the financial firm promoting the product cannot independently certify that they have recorded all trials, and controlled for selection bias, that investment ought to be presumed misleading. Instead, investors should purchase only those financial products where the firms have independently evaluated the Deflated Sharpe Ratio (Bailey and López de Prado [2014]), computed the Probability of Backtest Overfitting (Bailey et al. [2017]), or applied similar tests to control for SBuMT.

The academic community is aware of this financial research crisis (Bailey et al. [2014], Harvey et al. [2016]), and how financial firms are profiting from it. Preserving the *status quo* would not only be unethical, but outright fraudulent. The time for action is now.

APPENDIX

A.1. 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 U.S. dollars.

Avg Gini
Average of the daily Gini coefficients. The daily Gini coefficient is the ratio (i) and (ii), where: (i) is the area between the Lorenz curve and the line of equality, and (ii) 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

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

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 drawdown in percentage at the 95th percentile.

```
def computeDD_TuW(series,dollars=False):
    # compute series of drawdowns and the time under water associated with them
    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']] # hwm followed by a drawdown
    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%)

Time under water in years for the drawdown at the 95th percentile.

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.

FREQUENTLY ASKED QUESTIONS

1. **"Shouldn't a paper concern itself with the reason why a researcher tests a given 'strategy,' and the sharpness of her prior? Just disclosing the number of trials does not tell the full story, I think."**

We agree that the focus of a paper should be the theoretical justification for the prior that is being tested. However, having a convincing prior does not excuse scientific sloppiness or outright fraud. Every scientist must always reveal the extent of all trials involved in a discovery, so that referees can assess the probability that the claim is a false positive.

2. **"Suppose that I build a theory according to which a particular return-predicting factor (RPF) should be significant. The theory is true, and my one-trial experiment confirms it. There are 1,000 researchers and each one is allowed to guess only one RPF. One researcher guesses 'my' factor, but he has no idea why it should work. Even if we have both conducted only one test, it seems to me that my result is more interesting than his. How can this be captured by your approach?"**

The situation you describe concerns a true positive that someone found by accident. The purpose of our paper is not to prevent true positives (even if they come by luck), but to prevent the false positives that result from SBuMT. In any case, we agree with you that discoveries supported by theory should be preferred over purely empirical ones.

3. **"Isn't it true that a researcher may still find a false positive, even if he conducted a very small number of trials?"**

There are no infallible tests, with zero false positive probability. The goal of our method is not to reduce the false positive probability to zero. False discoveries will continue occur, at the rate set by that false positive probability. Our goal is to estimate that rate accurately.

4. **"If I have a very strange RPF, built with a weird combination of lags, variables, and exponents, and nothing else, it really smells of overfitting. But suppose that I arrive at exactly the same weird RPF from a theory that makes a very sharp prediction. All of the sudden, the same RPF becomes beautiful. How do we capture this?"**

If you have a theory, test it directly. Avoid engaging on a wide unconstrained search of alternative model specifications (backtest optimization). In that way, the number of clusters will be small (see Section 2.2), and the likelihood that your discovery is a false positive will remain low.

5. **"You rightly object to one researcher trying out a thousand permutations and reporting the good one. But, are we much better off if each researcher can only have a limited number of shots, or is taken seriously only if he reports a small number of shots?"**

There will always be researchers who find false positives, as predicted by the test's false positive probability (which is not zero). By not adjusting for SBuMT, journals have accepted false discoveries at a much higher rate than they expected. The great majority of false discoveries would have been prevented if journals had adjusted for the number of trials involved in a discovery. A small portion of false positives is inevitable, and our goal is to reduce that portion to the threshold accepted by the referee (the test's significance level). Once the full extent of the trials is taken into account, there is no reason to limit the "number of shots" given to researchers.

6. **"There are hundreds of thousands of researchers out there. Suppose the each of them controls for their own SBuMT. At a 5% false positives rate, there will be plenty of them submitting false discoveries to journals. How does your approach prevent that?"**

The problem you describe is real: Journals have a "publication bias" in the sense that they favor the publication of positive results. Authors who only found negative results may unselect themselves, hence journals are not exposed to all trials. Referees cannot control for trials that authors hide from them. One solution is that referees must require that authors run trials that other reasonable authors (who unselected themselves) would have attempted. Then, even if some trials are missing, the number of clusters will still be the same. The missing trials will be redundant, as they would have been folded onto clusters formed by the reported trials. A second solution is that journals share trials among themselves, in order to build a trials repository, like medical journals did with www.alltrials.net. As a side note, it would make sense for journals to publish negative results as well, or at least collect them in their databases. Negative results may not be monetizable, but they are useful from a research standpoint, as they help prevent false positives.

Fortunately, the question you raise is less relevant in the context of industrial research. Financial firms can legally enforce their right to record all trials used in selecting a strategy, and not only those that led to positives. There is no such thing as "publication bias" when a firm records all trials ever conducted, regardless of whether they led to a positive or not. This is a key advantage that industrial research has over academic research in finance. For further details see López de Prado [2017].

7. **"I am strongly in favor of showing the sensitivity of the results to changes in the parameters. Fine tuning smells of overfitting, but if the results are robust, then I can believe them more."**

We agree wholeheartedly. Authors must argue convincingly the robustness of their results. That will involve testing their models under alternative parameter values and specifications. Those

tests will be part of the trials set, and must be disclosed in accordance with rigorous scientific standards.

8. **"General relativity points to an uncomfortable degree of fine-tuning. Why is this more acceptable in physics but is less acceptable in finance?"**

Unlike in physics, finance does not have laboratories where theories can be tested independently and out-of-sample. Overfit physical theories can be debunked much more easily than in finance. That is why it is so critical in finance to prevent overfitting or selection bias in the first place. Once it has occurred, it may take many decades to gather the evidence needed invalidate the false claim.

9. **"Suppose that I have a hypothesis as to why an RPF should work. I try it and it does not. I look at my failure, analyze the data, and discover that the errors trace a parabola. Then, I deduce that my linearity assumption was too crude, and I must use quadratic terms. A lot of progress in understanding is achieved by 'playing' lovingly with the data. What constitutes data exploration and what constitutes a backtest? I think the boundary is porous."**

In your example, when you analyzed the data and recognized the pattern, you improved the strategy through understanding, not by sheer data-mining. Gaining that understanding means engaging in more trials. The objective is to gain understanding, while controlling for the probability that false positives occur under the guise of "understanding."

One important disclaimer is that a low false positive probability does not ensure success. It just tells us that the discovery is unlikely to be the outcome of trying random experiments and showing the best looking one. Most quantitative hedge funds engage in absurd backtest optimizations that invariably lead to backtest overfitting, false positives, losses and failure. Most of those failures would have been avoided if firms enforced scientific reporting standards such as the one presented in this paper.

EXHIBITS

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)	1000.00	1506.43
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
Efficient 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

Exhibit 1.1 – Performance statistics for the index and the selected strategy

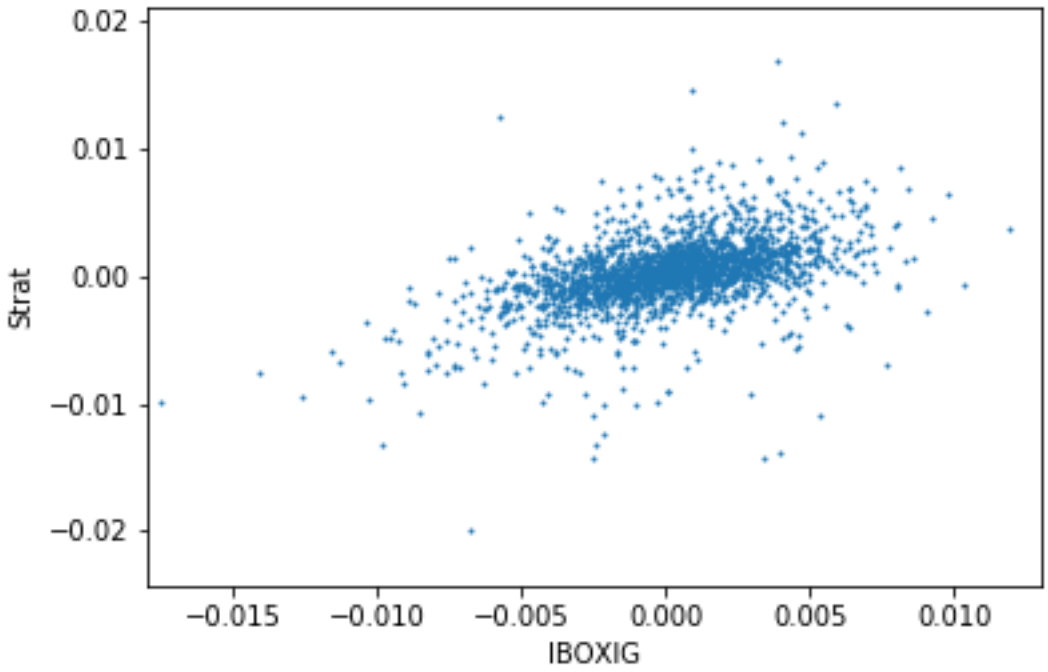

Exhibit 1.2 – Scatter-plot of iBoxxIG returns (x-axis) against strategy returns (y-axis)

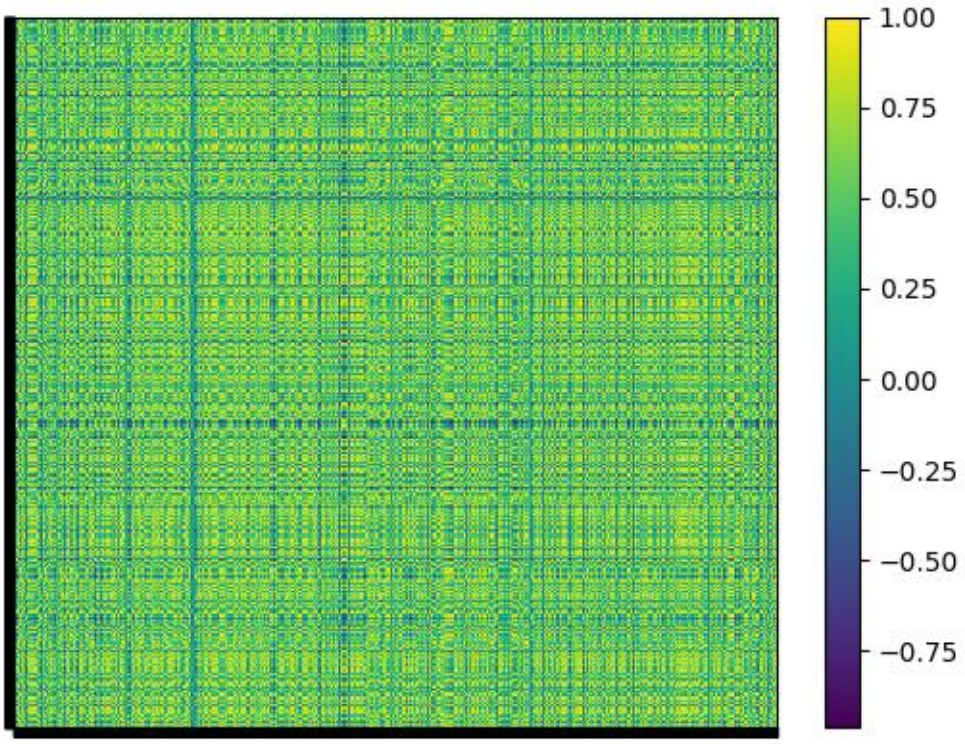

Exhibit 2 – Heatmap of the correlation matrix between the returns of all 6,385 trials

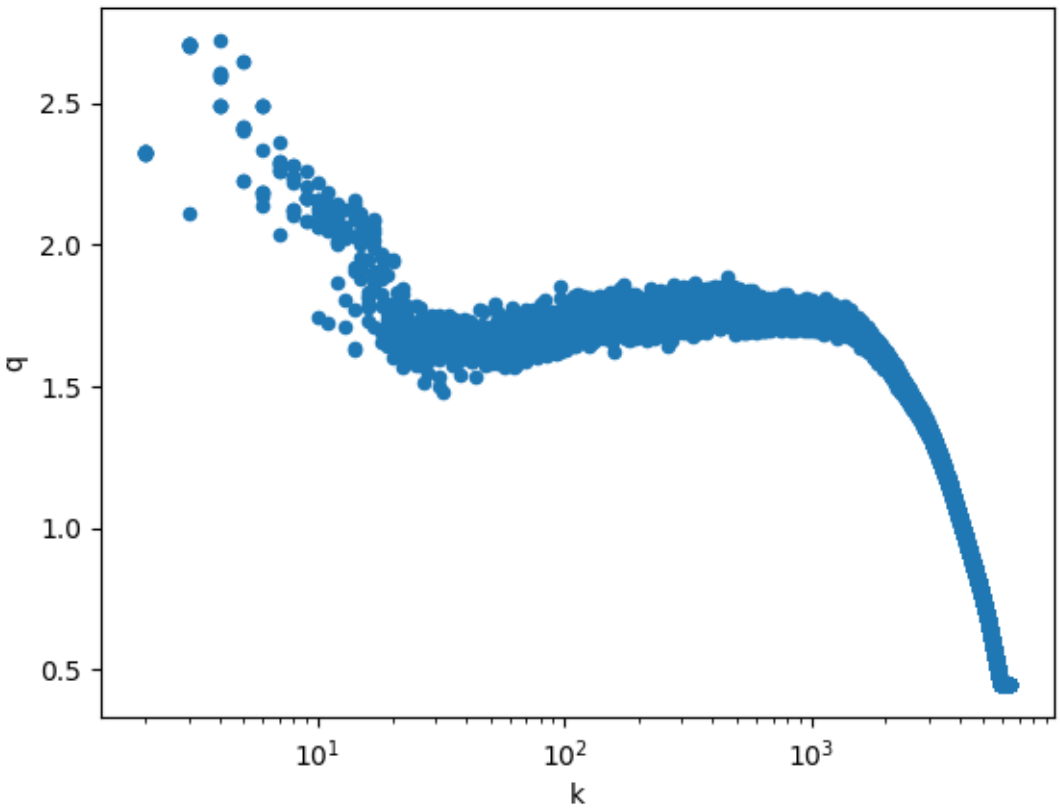

Exhibit 3 – Quality of clusters (y-axis, in log-scale) for a varying number of clusters (x-axis)

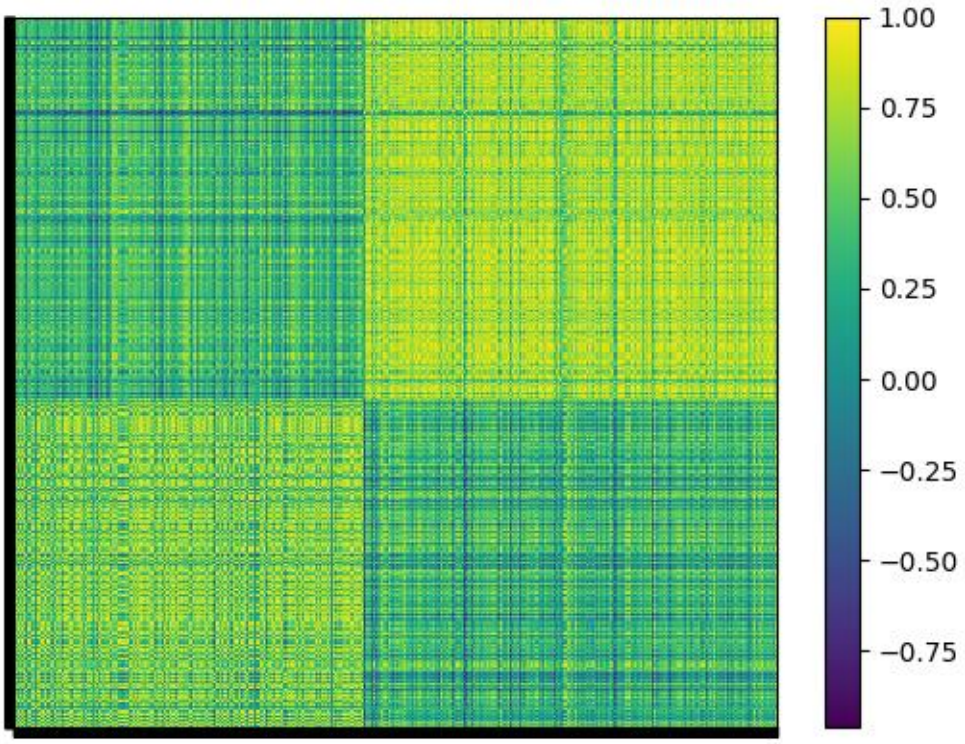

Exhibit 4.1 – Heatmap of the clustered correlation matrix, for k=2

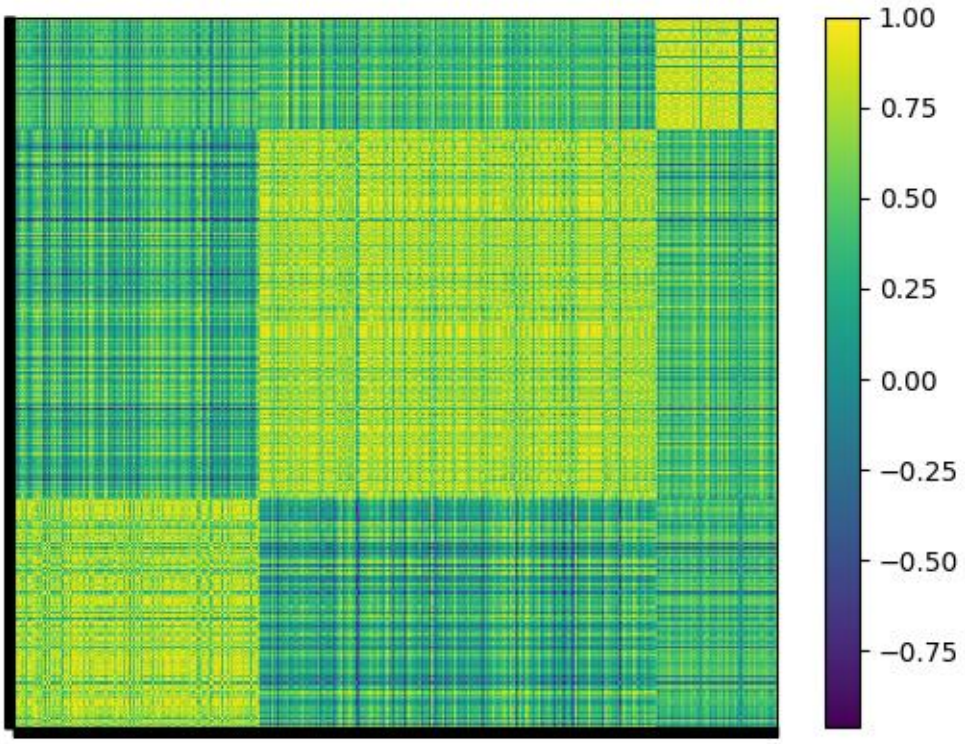

Exhibit 4.2 – Heatmap of the clustered correlation matrix, for k=3

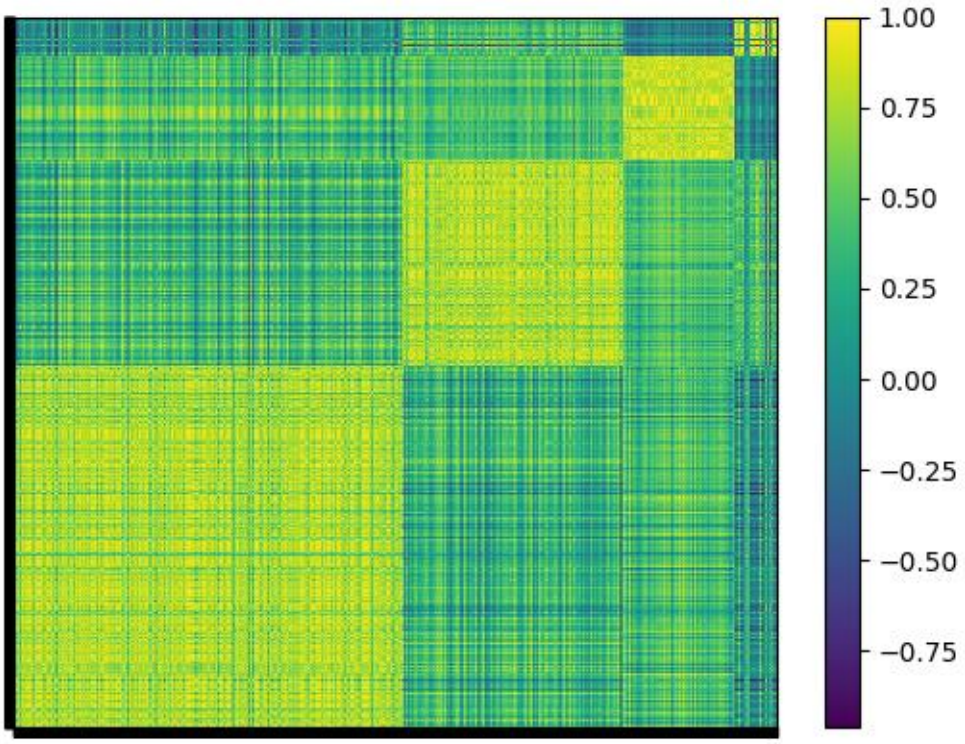

Exhibit 4.3 – Heatmap of the clustered correlation matrix, for k=4

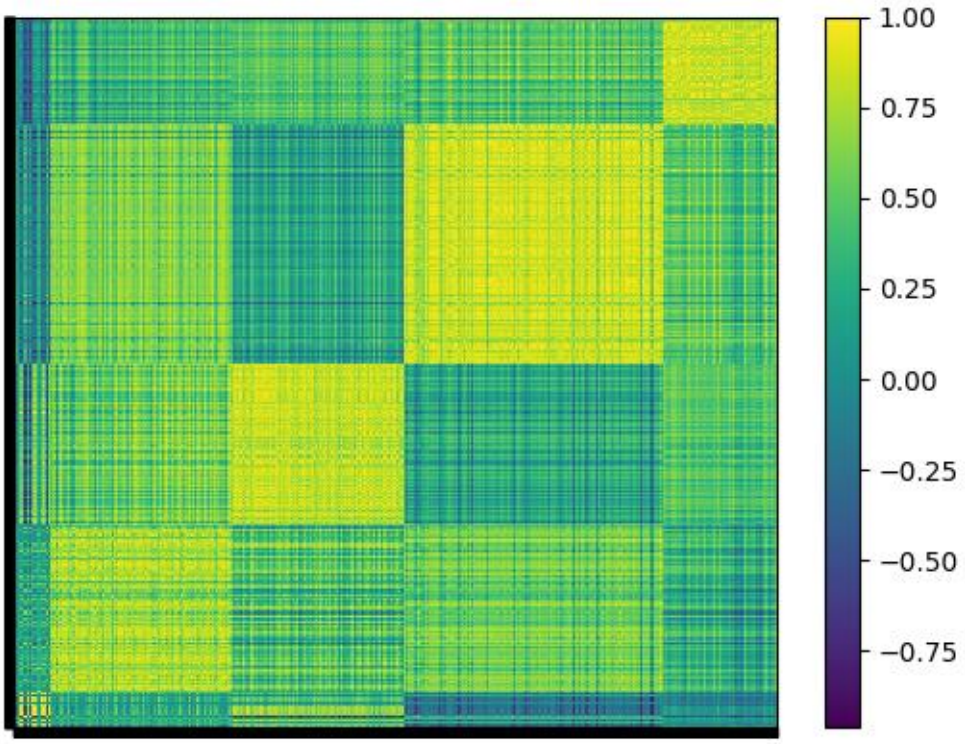

Exhibit 4.4 – Heatmap of the clustered correlation matrix, for k=5

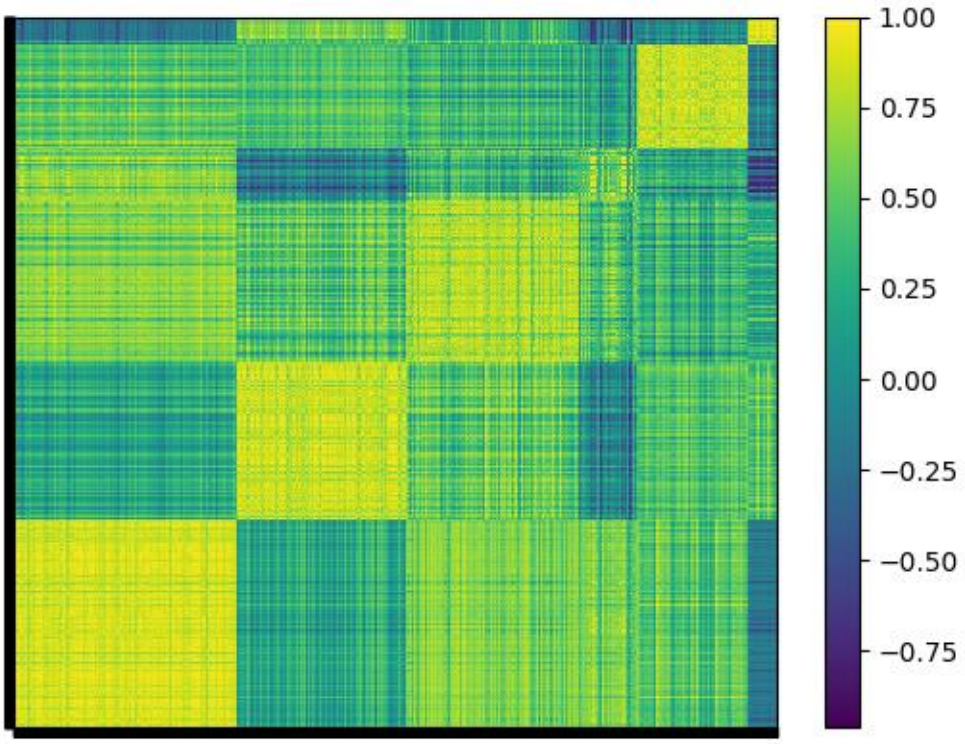

Exhibit 4.5 – Heatmap of the clustered correlation matrix, for k=6

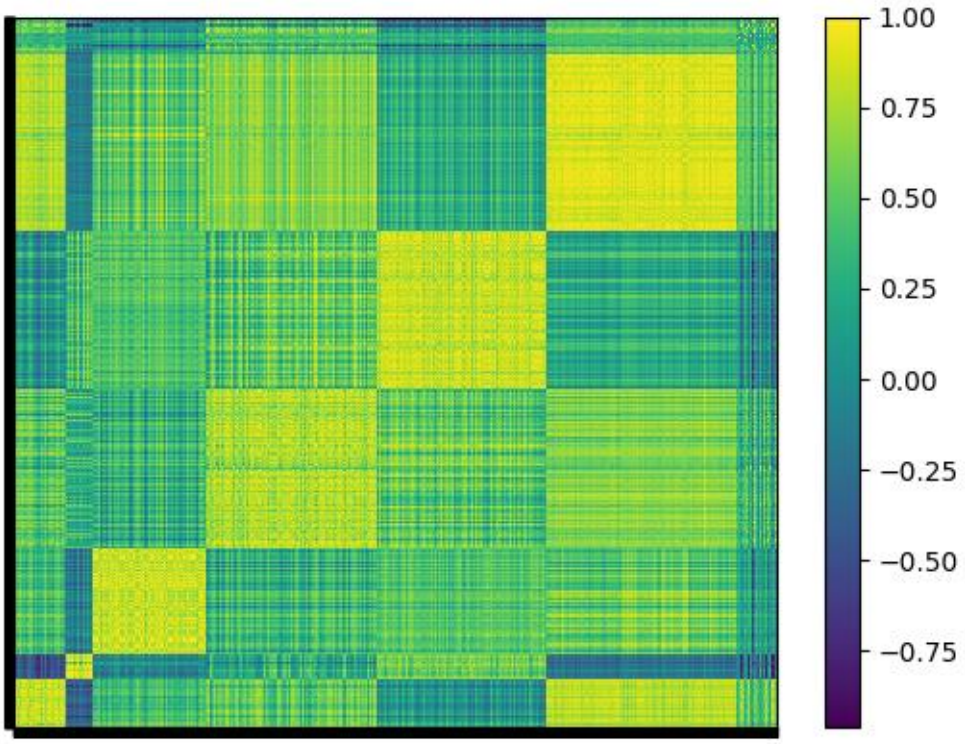

Exhibit 4.6 – Heatmap of the clustered correlation matrix, for k=7

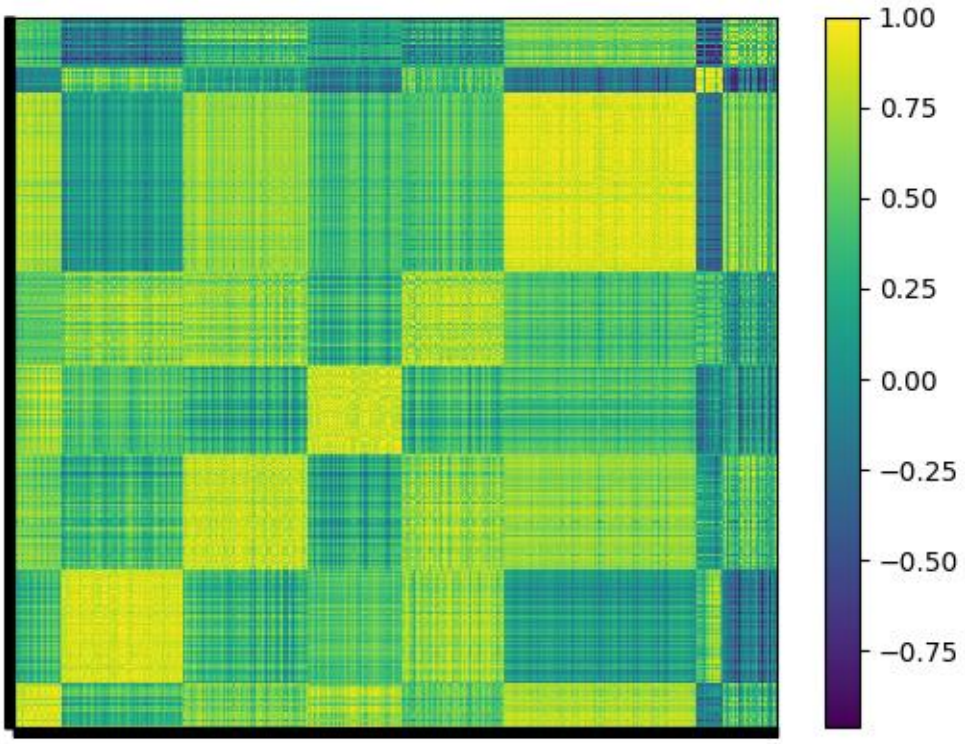

Exhibit 4.7 – Heatmap of the clustered correlation matrix, for k=8

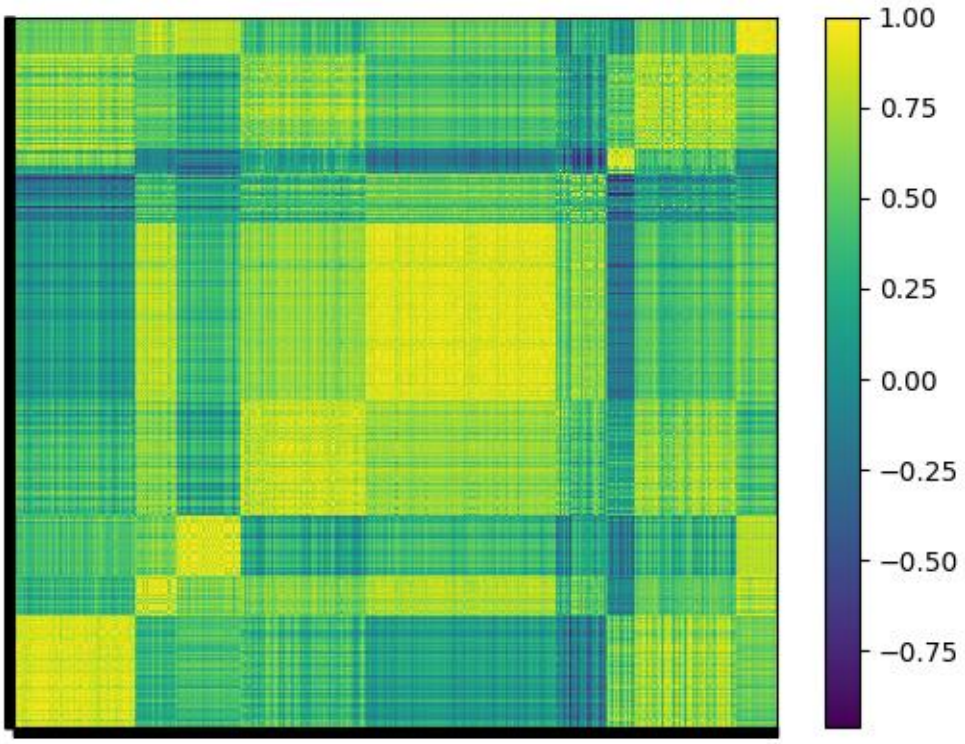

Exhibit 4.8 – Heatmap of the clustered correlation matrix, for k=9

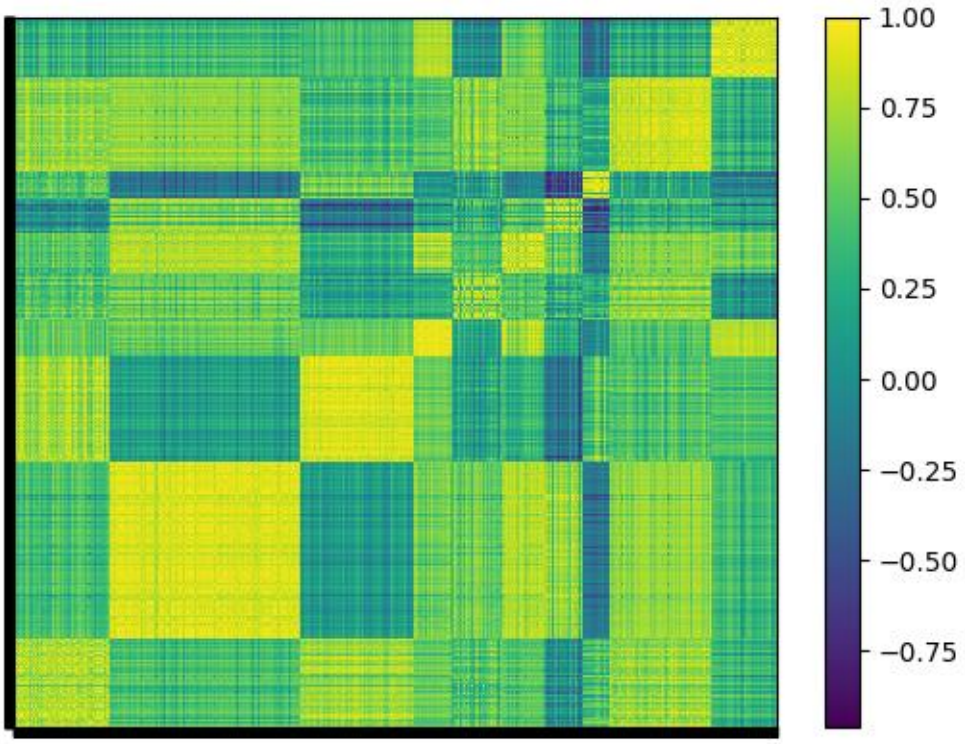

Exhibit 4.9 – Heatmap of the clustered correlation matrix, for k=10

Stats	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Strat Count	3265	1843	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	2172	2168	2174	2172
StartDt	2010-01-04	2010-01-04	2010-01-04	2010-01-04
EndDt	2018-05-01	2018-04-25	2018-05-03	2018-05-01
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

Exhibit 5 – Statistics computed on clusters' returns (k=4, q=2.7218)

Stats	Cluster 0	Cluster 1
Strat Count	2937	3448
aSR	1.7707	1.6023
SR	0.1096	0.0992
Skew	-0.5780	-0.3351
Kurt	6.5878	11.3212
T	2174	2172
StartDt	2010-01-04	2010-01-04
EndDt	2018-05-03	2018-05-01
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

Exhibit 6.1 – Statistics computed on clusters' returns (k=2, q=2.3274)

Stats	Cluster 0	Cluster 1	Cluster 2
Strat Count	2063	3329	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	2170	2172	2174
StartDt	2010-01-04	2010-01-04	2010-01-04
EndDt	2018-04-27	2018-05-01	2018-05-03
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**

Exhibit 6.2 – Statistics computed on clusters' returns (k=3, q=2.7068)

Stats	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Strat Count	317	1524	1434	2169	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	2172	2170	2168	2172	2174
StartDt	2010-01-04	2010-01-04	2010-01-04	2010-01-04	2010-01-04
EndDt	2018-05-01	2018-04-27	2018-04-25	2018-05-01	2018-05-03
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**

Exhibit 6.3 – Statistics computed on clusters' returns (k=5, q=2.6517)

Stats	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Strat Count	1873	1418	1447	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	2170	2168	2170	2172	2174	2170
StartDt	2010-01-04	2010-01-04	2010-01-04	2010-01-04	2010-01-04	2010-01-04
EndDt	2018-04-27	2018-04-25	2018-04-27	2018-05-01	2018-05-03	2018-04-27
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

Exhibit 6.4 – Statistics computed on clusters' returns (k=6, q=2.4919)

Stats	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Strat Count	443	232	940	1436	1418	1591	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	2170	2170	2174	2169	2168	2170	2172
StartDt	2010-01-04	2010-01-04	2010-01-04	2010-01-04	2010-01-04	2010-01-04	2010-01-04
EndDt	2018-04-27	2018-04-27	2018-05-03	2018-04-26	2018-04-25	2018-04-27	2018-05-01
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

Exhibit 6.5 – Statistics computed on clusters' returns (k=7, q=2.3650)

Stats	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Strat Count	411	1021	1037	794	846	1606	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	2170	2167	2169	2174	2168	2170	2170	2172
StartDt	2010-01-04	2010-01-05	2010-01-04	2010-01-04	2010-01-04	2010-01-04	2010-01-04	2010-01-04
EndDt	2018-04-27	2018-04-25	2018-04-26	2018-05-03	2018-04-25	2018-04-27	2018-04-27	2018-05-01
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

Exhibit 6.6 – Statistics computed on clusters' returns (k=8, q=2.2822)

Stats	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8
Strat Count	1021	352	536	1037	1593	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	2167	2170	2160	2169	2170	2172	2170	2168	2171
StartDt	2010-01-05	2010-01-04	2010-01-22	2010-01-04	2010-01-04	2010-01-04	2010-01-04	2010-01-04	2010-01-04
EndDt	2018-04-25	2018-04-27	2018-05-03	2018-04-26	2018-04-27	2018-05-01	2018-04-27	2018-04-25	2018-04-30
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

Exhibit 6.7 – Statistics computed on clusters' returns (k=9, q=2.2594)

Stats	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9
Strat Count	806	1596	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	2168	2170	2167	2171	2170	2170	2172	2170	2169	2160
StartDt	2010-01-04	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
EndDt	2018-04-25	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
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**

Exhibit 6.8 – Statistics computed on clusters' returns (k=10, q=2.2211)

REFERENCES

American Statistical Association (1999): "Ethical guidelines for statistical practice." Committee on Professional Ethics. Approved by the Board of Directors (August 7, 1999). Available at 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*, Vol. 61, No. 5, pp. 458–471. Available at http://ssrn.com/abstract=2308659

Bailey, D., J. Borwein, M. López de Prado, and J. Zhu (2017): "The Probability of Backtest Overfitting." *Journal of Computational Finance*, Vol. 20, No. 4, pp. 39-70. Available at http://ssrn.com/abstract=2326253

Bailey, D. and M. López de Prado (2012): "The Sharpe ratio efficient frontier." *Journal of Risk*, Vol. 15, No. 2, pp. 3–44. Available at https://ssrn.com/abstract=1821643

Bailey, D. and M. López de Prado (2014): "The deflated Sharpe ratio: Correcting for selection bias, backtest overfitting and non-normality." *Journal of Portfolio Management*, Vol. 40, No. 5, pp. 94-107. Available at http://jpm.iijournals.com/content/40/5/94

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

Harvey, C., Y. Liu and C. Zhu (2016): "…and the Cross-Section of Expected Returns." *Review of Financial Studies*, Vol. 29, No. 1, pp. 5-68. Available at https://ssrn.com/abstract=2249314

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

López de Prado, M. (2017): "Finance as an Industrial Science." *Journal of Portfolio Management*, Vol. 43, No. 4, pp. 5-9 (Summer). Available at http://jpm.iijournals.com/content/43/4/5

López de Prado, M. (2018): Advances in Financial Machine Learning. 1st edition, Wiley. https://www.amazon.com/dp/1119482089

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

Sala-i-Martin, X. (1997): "I just ran two million regressions." *American Economic Review*. Vol. 87, No. 2, May.

Sharpe, W. (1966): "Mutual Fund Performance", *Journal of Business*, Vol. 39, No. 1, pp. 119–138.

Sharpe, W. (1975): “Adjusting for Risk in Portfolio Performance Measurement", *Journal of Portfolio Management*, Vol. 1, No. 2, Winter, pp. 29-34.

Sharpe, W. (1994): “The Sharpe ratio”, *Journal of Portfolio Management*, Vol. 21, No. 1, Fall, pp. 49-58.

Szucs, D and J. Ioannidis (2017): “When Null Hypothesis Significance Testing Is Unsuitable for Research: A Reassessment.” *Frontiers in Human Neuroscience*, Vol. 11, Article 390. Available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5540883/