January 2017

Big Data in Investment Finance: A Cautionary Comment

Northfield Information Services Commentary

Dan diBartolomeo

As the availability of data on all manner of things has exploded in recent years, the investment industry has quickly embraced the concept of "big data" as the next way in which professional investment managers will gain advantage both over their peers, and over purportedly less sophisticated investors. Certainly there have been some investment entities (e.g. the Renaissance organization) that have achieved great success by purportedly recognizing patterns in the flow of events within, and exogenous to financial markets. Simple logic would suggest that the greater the scope of data available to analyze, the greater the number of useful patterns that might be discovered.

On the other hand, many organizations that have based strategies on the big data concept have dramatically underperformed expectations. For example, on the 9th of January 2017 Bloomberg News carried a story reporting the very poor 2016 performance of the various funds managed by the Blackrock "Scientific Active Equity" group. According to the article, "SAE is made up of more than 90 investment professionals, including 28 Ph.D.s and numerous data scientists." Clearly, having a staff that is well versed in the investment applications of "big data" is not a guarantee of success.

Unlike physical sciences, finance is "hand of Man, not hand of God." The geopolitical, economic and regulatory circumstances which impact financial markets change from day to day, sometimes in small and benign ways, but sometimes in profound ways. Those circumstances, known and unknown, exert a powerful impact on the usefulness of the assumption that statistical patterns when observed carry material meaning. It would seem useful to recall this quotation from the play *The Doctor's Dilemma*, written by George Bernard Shaw.

> *Even trained statisticians often fail to appreciate the extent to which statistics are vitiated by the unrecorded assumptions of their interpreters... It is easy to prove that the wearing of tall hats and the carrying of umbrellas enlarges the chest, prolongs life and confers comparative immunity from disease. A university degree, a daily bath, the owning of thirty pairs of trousers, a knowledge of Wagner's music, a pew in church, anything, in short, that implies more means and better nurture…. can be statistically palmed off as a magic spell conferring all sorts of privileges... The mathematician whose correlations would fill a Newton with admiration, may, in collecting and accepting data and drawing conclusions from them fall into quite crude errors by just such popular oversights as I have been describing.*

Big Data in Investment Finance: A Cautionary Comment

Beyond simply using “big data” techniques to uncover new information of potential usefulness, investment practitioners then routinely evaluate the strategic importance of such information through investment simulations known as “back-tests." Unfortunately, this is where the investment research process based on “big data” goes awry, as these simulation tests are deeply biased so as to offer “false positive” outcomes in a preponderance of cases.

The problem of “false positive bias” exists not only when big data is involved, and not only when addressing financial markets. It is routine that scientific and professional journals in all fields will not publish null results. If fifty teams of medical researchers study a given issue, and two studies claim to have found the same important result those two studies will be published, as they confirm each other. The fact that forty-eight teams of comparable researchers *did not find this result goes unpublished and unnoticed*.

In asset management strategies have been driven by either the human business judgment of fundamental investors, or the statistically driven results of quantitative investors. Traditionally, quant strategies have been based on hypotheses grounded in fundamental concepts. So too can many “big data” strategies be grounded in fundamental concepts. For example, if we use satellite imagery available on the Internet to count the number of vehicles in the parking yard of a large retail store, this could be very valuable in making a forecast of the store sales. We are simply using new data technology to do what could be done manually at greater effort (count the cars).

On the other hand, strategies could be based solely on the fact that we can use powerful computers to find patterns in large data sets that have hopefully gone unnoticed by others. Powerful biases associated with “data snooping” in finance have long been well recognized, as in Lo and Mackinlay (*Review of Financial Studies*, 1990). When we formulate strategies based purely on big data, the process of validating the usefulness of a strategy must dramatically be more rigorous than when large data sets are being employed to more efficiently implement what is otherwise a theoretically and intuitively appealing investment thesis.

Unfortunately, industry practices around the validation of investment strategies are woefully inadequate even for conventional methods, and even more so for data driven processes. In 1995, I did a lecture at a CFA event at Northwestern University, near Chicago. The title was a question “How to Blow a Back-Test?". Sadly, the answer was “believe the results." In the more than twenty years since that lecture, industry practices around “back-testing” have not changed much. Such simulations remain a staple activity of investment firms. Favorable back test results are an essential part of the marketing of new portfolio management strategies.

Conventional back-testing procedures almost always yield results that are statistically insignificant and have little real content. If anyone tells you to believe the results from a typical back test, they are *a liar, a fool or both.* In their seminal research study, Bailey, Borwein, de Prado, and Zhu (*Notices of the American Mathematical Society*, 2012) define a back-test as “a historical simulation of an algorithmic investment strategy." Their paper was aptly titled “Pseudo-Mathematics and Financial Charlatanism." They state: “A back test is useful and valid when the *correctly formulated expectation* of out of sample performance is equal to the in-sample simulated performance.”

Big Data in Investment Finance: A Cautionary Comment

Back-tests as carried out by investment practitioners almost never meet that standard of validity for a myriad of reasons. Stationarity assumptions and look ahead bias leading to invalid priors for expectations. Unrealistic test conditions (e.g. zero trading costs) further contribute to the upward biased outcomes. Most importantly, industry simulations are structurally limited to insufficient sample sizes. This flaw is then made worse by systematic application of "overfitting."

The whole concept of back-testing rests on the idea that history can be our guide to the future. There are extreme assumptions that the passage of time in financial markets is a stationary process. We assume that to a meaningful degree, future financial outcomes will be like the past. Kahn and Rudd (*Financial Analyst Journal*, 1995) show that this is a very weak assumption.

Almost all back-tests are look-ahead biased either in raw data or strategic concepts. Implicitly we are saying that "If I knew then what I know now, we would have performed well." While true on its face, basing strategies on this idea is very perilous. In his book, A Brief History of Time, the physicist Steven Hawking said that one of the proofs that time only goes forward is that "otherwise, we could invent a computer that would report tomorrow's stock prices." Look ahead bias in "data snooping" exercises can be substantially reduced by using databases that provide "point in time data" so you are seeing information (e.g. company financial statements) as they would have been at the time. Often investment analysts are using databases that reflect today's accounting standards at a particular firm, rather than what prevailed then.

Another real world aspect of financial markets that which is difficult to simulate is that everybody watches everybody else to see what's working, much like a fashion show. Simulations assume that a given financial market participant could have traded without any alteration in the course of past events. Given that every buyer must have a seller and vice versa this idea is just silly. It is doubly silly for large institutions that must do large transactions that will create obvious changes in the balance between supply and demand for a particular financial asset.

Most quantitative models for investment management rely on a series of parameters to formulate return expectations or do portfolio construction tasks. With the widespread availability of cheap computer power, there is an overwhelming urge to carry out many tests so as to find the best combination of the parameters. Carried to an extreme, this process is often automated as "machine learning." The greater the number of parameters and the extent of parameter calibration, the greater the length of the sample period needed to validate the process. Even if we make heroic assumptions about stationary processes, the required amount of historical experience often does not exist. Put simply, "overfitting" is rampant.

Conventional simulation tests have extremely low statistical power. Most tests simulate over the history that was actually experienced. There is rarely any testing of how a strategy would have done over the infinite number of other paths the evolution of history might have taken. There is no context to judge whether the sequence of events that we lived through was typical or improbable. Kahn (*Financial Analyst Journal*, 1997) offers some excellent examples of how low probability events are routinely misinterpreted.

Big Data in Investment Finance: A Cautionary Comment

When building and testing a quantitative model of investment returns it would be really good to know how many parameters you are looking for to explain what you have observed. An approach to this problem was offered in Bouchard, Laloux, Cizeau and Potters (*MMMAS*, 2000). In this study, they create long time series of simulated security returns using a random number generator. They then form the covariance matrix of these random series and do eigenvector decomposition (principle components analysis) to see how many apparently significant (but known to be spurious) eigenvectors they get. Depending on the number of simulated securities and the number of periods of simulated returns they often get many, when we know in advance the correct answer is always zero.

Even closer to the real world practice of modeling with big data is a recently updated study by Novy-Marx (http://rnm.simon.rochester.edu/research/MSES.pdf). This study extends the use of randomly generated "signals" to pick stocks. One can generate a large number of random signals and correlate each set of random signals to stock performance. We could then select several of the signals with the best fit in-sample. Next we could combine the selected signals into an "alpha" signal and scale to the expected distribution of returns. This "alpha" model would have a massively efficient fit to in-sample data while the correct expectation is zero predictive power out of sample. Since we don't think our real world alpha signals are totally random, the paper provides a way to analyze the extent to which out of sample results should be "deflated" to form realistic expectations of predictive power.

The aforementioned study by Bailey, Borwein, de Prado, and Zhu provides a very detailed analysis of the mathematics of "overfitting." Most investment models have multiple parameters. The greater the number of parameters, the larger the in-sample period must be to validate the process. A formula is provided showing that you need really long tests for the expectation out of sample performance to converge to the in sample simulated performance, even if we assume stationarity and distributional sufficiency. It should be noted that you can't get around this problem by simply reducing observation periods from months to weeks to days to get more observations. Many of the basic assumptions that we take for granted in statistical exercises, break down completely at the shorter end of the time scale, as described in diBartolomeo (*Professional Investor*, 2009). Alternatively, we can revise upward our requirements for statistical significance in considering simulated results to reflect the number of alternative parameter specifications we have tried, as in Harvey and Liu (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2345489, 2013).

A very useful reality check on expectations derived from simulations was provided by Sneddon (*Journal of Investing*, 2008). This technique provides a way to analytically forecast the long term performance (alpha, Sharpe ratio, etc.) of a strategy conditional on the forecasting power (IC) of the "signal." Given whatever you believe your information coefficient will be, the long term performance is predictable given certain assumptions. You have to assume your IC is positive but can vary over time, and that alpha signals decay with time so that the IC at a one time horizon may be different than at a longer horizon. Other relevant assumptions include that market impact portion of transaction costs is linear in trade size, and that the security covariance matrix used to estimate risk is accurate. This technique can be extended to the purpose of performance attribution of live results.

We are still in the early days of the application of "big data" to investment decision making. Much can be done to sensibly "supervise" the automated analytical processes that seek to find advantage in big data with respect to financial markets. If strategies based on the insights provided by big data are to succeed, the level of rigor being applied to evaluation of in-sample and simulated out of sample results must be far higher than today's common practice within the asset management industry, but can be addressed with more thorough but available techniques and tools.