OVERFITTING AND ITS IMPACT ON THE INVESTOR

MAN AHL ACADEMIC ADVISORY BOARD

May 2015

OVERVIEW

- Overfitting is when a model describes noise rather than signal. The model may have good performance on the data on which it was tested, but little or no predictive power on new data in the future. Overfitting can be described as finding patterns that aren't actually there
- There is a cost associated with overfitting – an overfitted strategy will underperform in the future
- We are biologically hard-wired to overfit: catching on to patterns can be an evolutionary advantage even if the pattern was not real. It is a phenomenon found not only among humans but also in other animals, such as pigeons
- Discretionary managers are as likely to overfit as quants, and may be less aware of it. Strategies based on theory are not immune to overfitting either
- Common methods to deal with overfitting include: using a range of models instead of relying on a single one (model confidence set), penalizing models for complexity, considering model stability under parameter variation, an independent review process, and monitoring the rejection rate of strategies through the review process
- Finance is not alone in facing the challenges posed by overfitting, it is a common phenomenon in science as well
- Data mining (the process of finding patterns in large data sets using computers) has become more popular in science, and most branches are using similar methods; in particle physics artificial data is used to optimize the analysis without overfitting, further methods include a rigorous review process and strict data partitioning
- Few branches of science have to deal with such inherent non-stationarity and feedback effects: in finance, the investment environment can change and a strategy that has worked in the past can stop working. Non-stationarity complicates the reproducibility of results
- A good way for an investor to see how a company deals with and reduces overfitting is by analyzing its culture. An environment with high tolerance for failure in research is likely to produce less overfitted results

CONTENT

www.ahl.com
www.man.com
www.oxford-man.ox.ac.uk

INTRODUCTION

The Man AHL Academic Advisory Board met in May 2015 to discuss overfitting and its impact on the investor.

The Board, whose members bring diverse perspectives and deep expertise, consists of:

- **Nick Barberis**
  Professor of Finance at the Yale School of Management
  – one of the world's leading experts in behavioural finance.
- **Campbell Harvey**
  Professor of Finance at the Fuqua School of Business at Duke University and Editor of the Journal of Finance from 2006-2012
  – a leading financial economist with a focus on the dynamics and pricing of risk.
- **Neil Shephard**
  Professor of Economics and of Statistics at Harvard University. He was the founding director of the Oxford-Man Institute of Quantitative Finance in Oxford and directed it from 2007-2011
  – one of the top theoretical and applied econometricians.

These distinguished academics* were joined by Sandy Rattray, CEO of Man AHL, Matthew Sargaison, Chief Investment Officer of Man AHL, Nick Granger, Head of Man AHL Dimension programme, Anthony Ledford, Chief Scientist of Man AHL, Jamil Baz, Chief Investment Strategist of Man GLG, Rob Furdak, CIO of Man Numeric, Shanta Puchtler, CIO and Director of Research of Man Numeric, and Marco – Andrea Buchmann, Quantitative Analyst at Man AHL (and previously researcher at CERN).

AGENDA

1. What is overfitting, and why is it important for the investor? What is the cost of overfitting?
2. Why does overfitting occur? Are we hard-wired to overfit, to find patterns where there is only noise?
3. Are there even strategies that are not overfitted? Are investment managers aware of overfitting?
4. How can you detect overfitting?
5. What approaches exist to assess or eliminate overfitting?
6. How is overfitting treated in other industries? What can finance learn from them?
7. What is the role of stationarity? Can we distinguish non-stationarity from overfitting once we go out-of-sample or into live trading?
8. What measures and methods can be used to reduce overfitting when looking for profitable strategies? How can the investor be confident that overfitting has been properly accounted for?

*The external members of the Man AHL Academic Advisory Board are compensated for their membership of the board.

THE DISCUSSION

1. Man AHL: What is overfitting, and why is it important for the investor? What is the cost of overfitting?

Cam Harvey (CH): Overfitting is when you propose an overly complicated model to explain something rather simple; it can also be that you found a simplified model that works only by chance. An overfitted strategy will likely underperform when faced with new data, be it the out-of-sample the analyst has not yet looked at, or when placed into live trading. This happens because the model does not adequately describe the effect but rather the noise in the data, so when you apply it to data it does not know, it becomes much less efficient. This type of data mining has been on the increase with today's easy availability of both financial data and computing power. There is just too much data mining going on now in finance!

Neil Shephard (NS): I think overfitting is not necessarily the right way to characterize the problem; you want *replication* when you build models. In a changing world, replication is very difficult. 'Overfitting' is an aspect of it, but the general problem is that you have to use the past to guide you towards making decisions for the future but you may be misled by the past. Past data. Past analysis of the data. Past glories. Misleading economic theory. All of this may confuse you, impacting how you formulate and validate your model. One should always keep in mind that the goal is to have a strategy that performs consistently.

Matthew Sargaison (MS): Twenty years ago we were facing different issues – today we have so much data that we have become more susceptible to overfitting.

Jamil Baz (JB): But data mining is not necessarily bad! There has been a major change in perspective in the past decades, with people starting out viewing data mining almost as a sin, but increasingly coming to appreciate it. One of the most positive aspects of data mining is that, if used correctly, one can stumble across things without needing the theoretical prior! Imagine for instance discovering the Black-Scholes equation just from options data instead of having to do the math…

NS: Science is clearly more open to data-based research than it was 20 years ago. Cancer research is a good example, there are a lot of studies with reproducible results that are not entirely understood yet. But we can learn from **[Cornfield et al, 1959]** and the more recent formalization by Ding and Vanderweele **[Ding and Vanderweele, 2014]**. Careful work will often demand very strong evidence before it is believed, as causal type inference can be misled by missing variables or false scientific understanding. Missing something, or overfitting in general, can lead to significant costs. An overfitted strategy, as Cam said, will underperform in the future, and I think one can summarize the cost of overfitting, and other effects like model misspecification, as being the difference between what you're selling and what the client gets.

Nick Barberis (NB): The cost can be far greater than the lack of reproducibility – investors would lose trust if too many quants overfitted. For instance, in some areas of psychology the lack of reproducibility has already reached crisis proportions, which seriously undermines confidence in published results.

2. Man AHL: Given the cost of overfitting, why does overfitting occur? Are we hard-wired to overfit, to find patterns where there is only noise?

NB: We do seem to be hard-wired to overfit. **[Tversky and Kahneman, 1971]** and **[Rabin, 2002]** have shown that people have a tendency to over-infer from small data samples; the gambler's fallacy is a good example. When we think of it from an evolutionary point of view, overfitting was probably helpful for our ancestors: if you heard a noise and assumed it was a predator then you might live to pass on your genes, while the person who didn't react might not be that fortunate. The preference for overfitting was therefore passed on. But it goes beyond offering an evolutionary advantage: we are also strongly motivated to find explanations for events so as to feel that we are in control. If we believe that Zeus causes earthquakes then we can appease Zeus to try and stop earthquakes. By 'detecting' and 'explaining' patterns in the data, people feel better, more in control. Confirmation bias just makes things worse. Once we have a hypothesis in mind, we are too accepting of further evidence confirming it and too closed to evidence against it.

Anthony Ledford (AL): There's a good example from behavioral psychology demonstrating that it is indeed not just a human trait – B.F. Skinner performed experiments with pigeons **[Skinner, 1947]**, which were placed in a cage and given food at random points in time. The pigeons then tried to catch on to a pattern, and repeated certain motions in an attempt to trigger food to be released even though the feeding was random. Overfitting doesn't seem to be limited to humans…

3. Man AHL: So if we are hard-wired to overfit, are there even strategies that are not overfitted? Are investment managers aware of overfitting?

Sandy Rattray (SR): I think the people who are most guilty of overfitting are discretionary managers, and as far as I can tell discretionary managers essentially only overfit. In other words, they are extremely fond of taking past single scenarios saying 'it happened in that scenario, it looks the same, therefore that is what we should do this time'.

JB: I agree, it relates, I think, to what philosophy calls Episteme and Techne. Episteme, or 'justified true belief' is when you have a repeatable experiment and then you can draw from statistical inference. Techne on the other hand is when the experiment doesn't lend itself to that kind of techniques. Examples of techne are some salient questions asked today to, and by, macro managers. Will QE work? What happens when the value of the stock market is multiplied by three and total debt to GDP increases simultaneously by 40 percentage points?

SR: But what discretionary managers will often do is say 'Well, Japan is the case example, we don't have many but we have one. So let's take Japan and what happened in Japan, ok, that's what will happen'.

JB: That's exactly right! The point is that if you have one data point, you extrapolate one point with one point which is a time-honored strategy in macro econometrics among discretionary managers, but if you don't have a data point then that's where things get interesting, that's where you need to be into techne instead of episteme, and that's where you need to conjecture about life...

SR: I think there is a bigger point which is that while discretionary managers are worse than quants, in terms of overfitting, actually everyone overfits. Are there any strategies that aren't overfitted?

Rob Furdak (RF): What about strategies that are derived from theory?

CH: Theory is flexible, so theory alone is not enough, plus sometimes theory is also overfitted.

NB: I don't think strategies based on theory are 'immune' either – theory can be helpful, but it does not protect you from overfitting. Nonetheless, trying to formalize an argument and doing the math you can sometimes already see the logic not working out, so in that sense theory is a good defense against overfitting, but it is far from foolproof, especially since theory is based on assumptions and parameters that can easily be adjusted.

4. Man AHL: How can you detect overfitting?

CH: There are a couple of obvious red flags: a strategy can for instance not make any economic sense, it can be counter-intuitive. It can also contain an unreasonable number of parameters to explain something rather simple, parameters that do not make sense. At the same time, determining whether a strategy was overfitted is difficult, as the reasons why a strategy can stop working are not just limited to overfitting – there can also be a structural change or an inefficiency people have arbitraged away.

NS: In theory this is not a very difficult problem. Bayes' theorem tells you how to test this, you look at evidence conditional on the model, and you can penalize models based on complexity.

JB: There is also an implicit survivorship bias, if you have a database of price or other economic data available then you are automatically looking at a country which has done rather well.

NS: And the world keeps on changing, which poses the most fundamental problem for replication.

AL: That is why finance is different from, for example, particle physics or genomics. There is feedback – market change drives sentiment and regulation which then changes the market. A lot of strategies capture small effects, they only 'barely work', so it's very difficult to assess whether they're still capturing an effect.

Shanta Puchtler (SP): What if quantitative investment managers artificially add noise to the input, or output, dataset, and analysts then come up with a strategy? As a part of the tests, aside from looking at the out-of-sample, one would then remove the artificial noise, and if the effect is real, removing the noise should improve the performance. On a related but somewhat different note, one could even think of a robustness test: what happens if outliers are removed or trimmed, does it still perform well?

Marco-Andrea Buchmann (MAB): Something along these lines is common practice in particle physics in the context of precision measurements. The researchers can test their algorithm and extract the result on a deliberately 'modified' dataset – once they are confident that their algorithm works they look at the unmodified data, they run the same algorithm again and report the final result. So this goes in the direction of Shanta's idea.

NS: It is important to keep in mind that Type I & Type II errors weren't designed to meet the needs of finance, so the academic literature should get more skeptical as more and more papers are published on the same dataset using these blunt tools. One should therefore have different methods to decide on the success of research methods.

5. Man AHL: What approaches exist to assess or eliminate overfitting?

NS: Instead of finding one excellent model, one can find a set of models – that's the idea of the 'model confidence set' (**[White, 2000]** and **[Hansen et al., 2011]**). This is the set of models that are statistically competitive based on criteria like a drawdown measure, a risk measure, and so on. Secondly, you can penalize models for complexity, e.g. using marginal likelihood. Furthermore, you can use model averaging instead of searching for a single model. Finally, in data science, there's been a lot of movement away from discussing overfitting towards considering regularization: when you have a lot of regression coefficients you should shrink them towards your prior, for instance using the LASSO method where you can shrink parameters exactly to zero. An example of the use of LASSO type arguments is in portfolio allocation, where you can impose that you short less than 25%. Outside of finance genetics is also a good example, they look at regularization when they face 10,000s of genes but a sample of 1,000 people.

CH: A common approach is to look at how slightly different models perform, i.e. you modify the model parameters and look at the corresponding 'heat maps', and assess how performance changes when varying the parameters. If your parameter choice is an isolated point sticking out, then this is a red flag and suggests the parameter choice was a result of data mining.

NS: Which is the exact opposite of what you are looking for in statistics, where you want things to be highly pointed, so that there is a perfect solution and the rest is just flat. Statistical thinking as it is usually stated is not ideal for this kind of problem though.

JB: There is also an interesting paper **[Ioannidis, 2005]** where the author talks about the number of true relationships versus the number of 'no relationships', among all tested in the field. This number is very small in many areas, leading to most results being false, in fact the smaller the number of true relationships compared to all relationships, the more likely you are to find a relationship that isn't a 'true' one. What do you think this ratio is in finance? I would say it's very low.

Nick Granger (NG): I don't think the notion of 'true relationships' is even applicable to finance due to extreme non-stationarity and feedback effects, but the signal to noise ratio in finance is definitely very low.

JB: Yes, so it's important to monitor our strategies. There's the concept of the rejection rate, so basically monitoring how many strategies are proposed originally, and how far they make it through the process. How many don't survive the initial research phase, how many are killed during review, and how many don't make it past test trading – monitoring this rate can be an important asset.

CH: AHL looked at the probability of overfitting **[Bailey et al, 2015]**, what became of it?

MS: Yes we have looked at that. The assumption of mean reversion is a bit strong, but we have used the ideas to enforce data partitioning with in-samples and out-of-samples more strongly. You now have to write down the data partitioning and the methodology and expectations.

SR: Maybe we could use fake data, and see how the strategy performs?

MAB: Something similar is done in particle physics; I've previously mentioned deliberately modified datasets, but we also generate fake data before analyzing any real data we get from a collider. This serves a range of purposes, the most important ones are of course understanding a new machine and studying it but also understanding the signal, seeing how we can improve on it assuming that it exists, and therefore improve the signal to noise ratio. The idea is to prepare a well-designed hypothesis test without optimizing the parameters to noise in the real dataset.

6. Man AHL: How is overfitting treated in other industries? What can finance learn from them?

MS: It is surprising how often the pharmaceutical industry has been cited by other competitors as a best practice arena given some of the very poor results with hidden datasets and bad practices ... but what can we learn from other industries?

CH: The types of issues vary across different industries, and the pharma industry is indeed under fire. There is an organization called alltrials.net that has nearly 100,000 signatures in a petition that demands that the results of all clinical trials be made public. What does it mean if a company conducts 19 trials and each one fails, they keep the result secret, and then report the results of the 20th trial with supposedly 'significant' results? Hidden tests are a huge problem but not just for pharma but finance too. An investment manager sees the results of a junior researcher but may not know how many strategies were tried. Even worse, a client might be presented with a very select result not knowing that hundreds of trials were conducted to get to the one she is seeing. It is crucial that investment managers keep a record of what was tried.

MS: This is something we do internally at AHL, we do keep a record of the things we try. At the industry level, on the other hand, this will never happen. One major difference between pharmaceutical companies and finance is probably that drug companies stand to benefit from false positives as their drug appears to work and they can monetize it. We on the other hand would be trading something that is just randomness.

NG: It may be worse than that: in finance, with the exception of transaction cost, a false discovery is likely to make you zero so while it's not great it's not that horrible either. In medicine if you do something wrong it could make things much worse than zero.

NS: In a somewhat related field, genetics, a vast number of hypotheses are tested, and the procedure is to look at vast numbers of relationships. Once a number of candidates has been found, the focus then shifts to these alone to analyze and understand them in detail – data mining is used as a type of exploratory analysis.

MAB: This is common in many branches of science, and they often find themselves in a similar situation as finance. Due to funding pressure, scientists need to publish, and null results in some areas, such as medicine, don't get published, so non-null results appear – that's almost a recipe for overfitting. In particle physics journals do publish null results, but people are very much aware of potential overfitting. CERN takes a range of measures, including artificial datasets, control regions, strict enforcement of in-sample and out-of-sample, and a rigorous review processes with other teams replicating the results. All of these really help with overfitting and maintaining high quality research.

NS: Campbell, you were the editor of the Journal of Finance for a while, were there any replication studies in finance?

CH: Replication studies are rarely published in the top journals. Why? It is true that the author may have looked at many variables to get the relation, but once you have the relation, the data are available for anybody to replicate the basic idea very quickly and cheaply. Most of the replication in finance is conducted by Ph.D. student class projects or by investment managers checking to see if the academic findings might be useful in the practice of asset management. In medicine, replication is expensive but important. If a medicine is going public, it is important to have as many independent tests as possible – to rule out the possibility that the findings of one test were just a fluke. In psychology, replication is very expensive, too. However, unlike in medicine, there is no incentive to replicate the studies – mainly because it is not a matter of life and death. It is therefore much more difficult in psychology to separate the true findings from the lucky.

NB: This issue has reached crisis proportions in some areas of psychology and it is really undermining confidence. They are focusing attention on ideas that are similar to those in other fields, such as the p-curve and p-hacking. Another idea is to record in advance what experiment you are going to run and so on. It's striking to me how right now, this issue has really come to a head in many different fields.

JB: It seems to me that unlike medicine or physics, we are dealing in finance with first order random walks that may arise from rational expectations. On top of that there's a feedback mechanism between the model and the agents, which means that the burden of proof is that much harder in finance than in other areas.

7. Man AHL: What is the role of stationarity? Can we distinguish non-stationarity from overfitting once we go out-of-sample or into live trading?

NB: There are reasons for non-stationarity and the mechanisms can be tested, for instance if you believe that the effect was arbitraged away you can look for the effect in markets that are harder to trade. Of if you believe there was a structural break then you can test for that as well.

CH: It is important to note though that typically one does not instantly switch from one regime to an other, so it may not be that easy to detect.

AL: There are ways to disentangle that: If you have a range of strategies, and you see the same breaking point across all of them simultaneously, then a regime shift is likely. If, however, the breaking point only appears for one strategy then it is more likely due to overfitting.

MS: Unfortunately most strategies have low Sharpe Ratios so even if you believed that there was a structural change you'd have to wait a while for the underperformance to become significant.

NG: I think that's something that's fundamentally different between physics and finance; in finance, the probability of a regime change is always greater than zero. In most branches of science stationarity assumptions are much less problematic.

8. Man AHL: What measures and methods can be used to reduce overfitting when looking for profitable strategies? How can the investor be confident that overfitting has been properly accounted for?

MS: What would you recommend, Cam, in terms of the due diligence process for an investor, to come into Numeric or AHL, or anybody? Everybody will say that they have the best execution, the best research team, the best shiny building, what do you actually need to look for and ask?

CH: I had a job in the past working for a major US pension fund, doing due scientific diligence on prominent asset managers. One thing that I looked for that was hard to measure quantitatively was an assessment of the research culture within the firm. To minimize overfitting, you can impose certain safeguards, e.g. the company tracks the number of times that researchers access the data or run different models. However, researchers can often bypass these safeguards if they want to mine the data. I always look for the 'culture of failure' at the firm. Suppose two researchers propose two ideas and both of the ideas are deemed to be high quality. Resources are allocated to the investigation. Both of the researchers do high quality work. The first researcher's hypothesis is supported by the tests. The second researcher's idea fails to get support in the data. Again, to emphasize, both researchers proposed high quality ideas and both executed tests with a high degree of competence. The second researcher should not be punished just because the finding was negative. This researcher did good work. If there are punitive consequences to the second researcher, this will lead to potentially extreme data mining. The researchers know they must show something 'significant' or there are negative consequences. Of course, these 'significant' strategies are doomed to fail in client trading. Therefore, it is important to have a culture that rewards research failure given the work was high quality.

SR: People generally don't ask very much about overfitting. And secondly, I don't think we've made a particularly convincing case that if the t-stat is 3 instead of 2 and we use the out-of sample in a clear way that the investor's returns will be better. And I'm not sure it's as convincing as we wish it was ... and that's a difficult thing. I'd be delighted if someone told me I was wrong!

CH: It seems clear that the investor should demand some sort of explanation of how the firm deals with these issues. If the investment manager has not thought through these issues, that is a red flag for me and I would avoid investing with them. Of course, there is a tradeoff, as Sandy mentions. If the hurdle for declaring a strategy significant is too high, this will greatly reduce the chance that money is allocated to false strategies that were just a result of data mining. However, importantly, it will also cause the company to miss some strategies that are true money making strategies. It is up to management to work out the balance. However, it is impossible to even approach this problem without thinking deeply about the firm's research process.

MS: A key message we try to give people is that all strategies are first traded with the firm's capital.

NG: The interesting thing is that the 2-3 months we would typically test trade a new model is not enough time to demonstrate any kind of statistical significance around the performance. However there have been a number of occasions where this process has caught 'bad' strategies. This could uncover a mistake in the research or the implementation. A model immediately going into a large drawdown would certainly be a warning sign. For models with high expected Sharpe ratios we can be more certain: for example a 20% drawdown over the first three months would certainly look inconsistent with an assumed SR of 3.

NS: If I'm trading with a Sharpe of one, and I think my overfitting/model error consensus Sharpe is ±0.2, then as a professional I would be surprised if I didn't achieve 0.8 in trading, but I would also be surprised if I surpassed a Sharpe ratio of 1.2. It would be useful to develop a standard for quantifying the risk of overfitting or more generally model error.

APPENDIX: SUMMARY

The issue of overfitting can manifest itself in two different ways: either by testing multiple possibilities at once ('multiple testing problem') or by using an overly complicated model to describe something simple ('backtest overfitting'). The characteristic property of overfitting is that the resulting pattern may appear to have high significance (in the context of finance, good performance) when in reality the pattern has arisen purely by chance, and describes the noise in the data rather than a profound property. This can lead to significant cost: a strategy describing noise rather than a fundamental property of the data will likely fail in the future, and investing in such a strategy will likely lead to underperformance in the future.

The literature on the topic has expanded considerably in the last decades; the pioneers in the field were John W. Tukey **[Tukey, 1951]** and Henry Scheffé **[Scheffé, 1959]**, but a very broad range of methods have been proposed since their original papers.

Finance is not alone in facing the issue of overfitting – numerous branches of science have faced similar challenges. CERN, for instance, has had to deal with the issue for the discovery of the Higgs boson **[CMS Collaboration, 2012] [ATLAS Collaboration, 2012]**, and NASA also has to deal with the effect **[Vittels and Gross, 2011] [Gross and Vitells, 2010]**. Further examples include medicine and biology, where questions of multiple testing routinely arise when considering multiple possible drugs. The effect is particularly pronounced in genetics, where linking complex diseases to environmental factors and multiple genes can quickly lead to thousands of tests.

There is no common approach to deal with overfitting, each branch of science has dealt with the issue in its own way; in finance, on the other hand, correcting for the effect is still far from standard. Issues around overfitting in finance are complicated by a combination of non-stationarity, feedback effects, and very low signal to noise ratios. While these issues are present to a greater or lesser extent in other fields, their pervasiveness in finance means that on top of applying statistical tests, more wide-ranging approaches should also be employed to take e.g. research culture and incentive structure into account.

APPENDIX: GLOSSARY

In sample: refers to the data and time periods used to fit, or build a model for trading. If the model has been estimated over some, but not all, available data, the remaining data is held back 'out of sample' and can potentially be used to evaluate the quality of the model.

Out of Sample: refers to the performance of trading a model away from the fitted back-test. Typically this means both the simulated performance using data that was held back from the original fitting and also the performance once the model is actually being used to trade with money.

Data Partitioning: the practice of dividing the dataset into two sub-datasets, one (the in-sample) for use in the analysis, and a second (the out-of-sample) for validation (see above).

Data Mining: the process of finding patterns in large data sets using computational analysis.

Backtesting: the process of testing a given strategy or model using historical data. The goal is to estimate the performance of the strategy or model to determine how it would have performed had it been used in the past.

Stationarity: refers to a process whose properties do not change with time. In the context of quantitative investment models, it typically refers to predictive relationships that do not change through time.

Type I error: refers to a 'false positive' model discovery, e.g. claiming that there is a predictive statistical relationship in a new model when in fact there is none.

Type II error: refers to a 'false negative', e.g. rejecting a proposed model when in reality there is a predictive statistical relationship in the data.

REFERENCES

ATLAS Collaboration, 2012
The ATLAS Collaboration. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. Phys.Lett., B716:1–29, 2012. doi:10.1016/j.physletb.2012.08.020.

Bailey et al, 2015
Bailey, David H., Borwein, Jonathan M., Lopez de Prado, Marcos, and Zhu, Qiji Jim. *The Probability of Backtest Overfitting*. Journal of Computational Finance (Risk Journals), Forthcoming, 2015.

CMS Collaboration, 2012
The CMS Collaboration. *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC*. Phys.Lett., B716:30–61, 2012. doi:10.1016/j.physletb.2012.08.021.

Cornfield et al, 1959
Cornfield, Jerome, Haenszel, William, Hammond, E. Cuyler, Lilienfeld, Abraham M., Shimkin, Michael B., Wyner, Ernst L. *Smoking and lung cancer: Recent evidence and a discussion of some questions*. Jounal of the National Cancer Institute, 22, 173-203, 1959.

Ding and Vanderweele, 2014
Ding, Peng and Vanderweele, Tyle J. *Generalized Cornfield conditions for the risk difference*. Biometrika, 101, 4, 971-977, 2014.

Gross and Vitells, 2010
Gross, E. and Vitells, O. Trial factors or the look elsewhere effect in high energy physics. Eur.Phys.J., C70:525–530, 2010. doi:10.1140/epjc/s10052-010-1470-8.

Hansen et al., 2011
Hansen, Peter R, Lunde, Asger, and Nason, James M. *The Model Confidence Set*. Econometrica, Vol. 79, No. 2, 453-497, 2011.

Harvey et al., 2014
Harvey, Campbell R., Liu, Yan. *Evaluating Trading Strategies*. The Journal of Portfolio Management, JPM 40, Vol. 40, No. 5, 108-118, 2014.

Ioannidis, 2005
Ioannidis, John P.A. *Why Most Published Research Findings Are False*. PLoS Medicine, 2 (8), e123, 2005.

Rabin, 2002
Rabin, Matthew. *Inference by Believers in the Law of Small Numbers*. Quarterly Journal of Economics, 117 (3): 775-816, 2002.

Scheffé, 1959
Scheffé, H. *The Analysis of Variance*. Wiley, New York, 1959

Skinner, 1947
Skinner, B.F. *'Superstition' in the pigeon*. Journal of Experimental Psychology, 38:168-172, 1947.

Tukey, 1951
Tukey, J. W. Reminder sheets for '*Discussion of paper on multiple comparisons by Henry Scheffe*'. In '*The Collected Works of John W. Tukey VIII. Multiple Comparisons: 1948-1983*' 469-475. Chapman and Hall, New York. 1951.

Tversky, Kahneman, 1971
Tversky, Amos and Kahneman, Daniel. *Belief in the law of small numbers*. Psychological Bulletin 76 (2): 105–110, 1971.

Vitells and Gross, 2011
Vitells, O. and Gross, E. *Estimating the significance of a signal in a multi-dimensional search*. Astroparticle Physics, 35:230–234, 2011. doi:10.1016/j.astropartphys.2011.08.005.

White, 2000
White, Halbert. *A Reality Check For Data Snooping*. Econometrica, Vol. 68, No. 5, 1097-1126, 2000.