STANDARDS BOARD FOR ALTERNATIVE INVESTMENTS

Backtesting

Key Questions for Investors to Ask

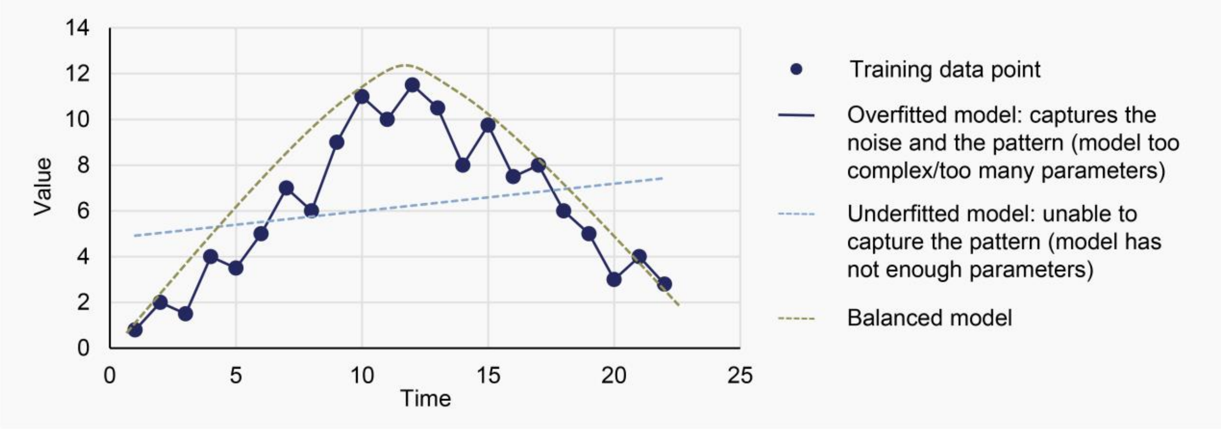
SBAi Toolbox

Introduction

Some investment strategies follow a mechanistic rule-based implementation which is fixed in time, for example in the area of alternative risk premia strategies (aka dynamic beta). These strategies are usually developed using backtesting (historical simulations), which applies the strategy to historic data to assess how the strategy would have performed in the past. The output of such a backtest is a time series of profits and losses for the strategy, which can then be summarised by risk adjusted return metrics (such as Sharpe Ratio). Also, the correlation with the return streams of other asset classes can be calculated.

A key challenge with using backtesting in strategy development is "statistical overfitting bias": This arises when a strategy is fitted too closely to the underlying historic dataset but might not work well in the future.

Overfitting vs. underfitting bias



In particular in situations where computers can test thousands or millions of different strategy configurations on a given sample dataset to find the optimum approach (e.g. highest risk adjusted return), it is very likely that the chosen strategy configuration will be overfitted, but with no superior predictive power in the future. Similar issues can arise where deep learning techniques/neural networks are employed (in conjunction with classic Fama French type factor models) to extract "deep factors" hidden by unexplained alphas of the benchmark model.

Therefore, managers and broker dealers using backtesting in strategy development need to develop frameworks and methods to vet strategies to prevent backtest overfit, to ensure that only those strategies that are deemed to have significant predictive power going forward are being implemented. From an investor's perspective, in situations where backtests are presented for a given investment strategy, it is important to assess the validity of the backtesting results, including the underlying assumptions and approach to identify potential overfitting bias.

The SBAI Toolbox is an additional aid to complement the SBAI's standard-setting activities. While alternative investment fund managers sign up to the Alternative Investment Standards on a comply-or-explain basis, the SBAI Toolbox materials serve as a guide only and are not formally part of the Standards or a prescriptive template.

Framework to assess backtesting results (including key questions for investors to ask)

1: Backtesting results ≠ past performance

Backtesting results constitute hypothetical performance information, not actual past performance

- Managers/broker dealers have to provide clear disclosure explaining how the backtesting results were derived, that the result is not the performance of any actual account and that it is not a guarantee of future results[1]
- Investors need to distinguish between historical information and backtesting results – **they cannot be compared**
- For funds with an actual performance history, managers should be careful to clearly delineate backtest results from actual performance in both graphical and textual presentations. After a fund has a live performance history of a sufficiently long duration, managers should consider whether backtest results should be presented in marketing materials at all.

  Observation: "When evaluating a trading strategy, it is routine to discount the Sharpe ratio from a historical backtest. The reason is simple: there is inevitable data mining by both the researcher and by other researchers in the past."[2]

2: Detecting statistical overfitting bias

Assessing whether the strategy configuration has been fitted too closely to the sample data

- Has the strategy been tested on out-of-sample data? (applying the strategy to data that has not been used in the initial backtesting phase)
- What backtesting techniques have been employed to avoid overfitting (train test split, multiple train text split, rolling window approach…)?
- How many trials have been undertaken to come up with the strategy?[3] (incl. looking at minimum backtest length for a given number of trials)
- Calculate metrics such as probability of backtest overfitting, performance degradation and probability of loss, stochastic dominance, etc[4]
- Assess performance (and risk) impact of strategy enhancements ("naked" versus enhanced strategy back-testing performance) to assess the risk of overfitting due to excessive complexity or parameters too closely fitted to the specific sample set
- What adjustments to the backtest are undertaken (e.g. applying a haircut to Sharpe Ratios or introducing a / "profit hurdle" for strategies (to be deemed "significant")[5]
- What governance arrangements are in place to prevent statistical overfitting (e.g. "Index Validation Committee")?

3: Assessing underlying assumptions

It is important that the backtest be run using realistic "real life" trading assumptions

[1] See https://blogs.thomsonreuters.com/answerson/back-tested-performance-misleading-not-off-limits/ on advertising actual vs. model vs. backtested performance; specific SEC advertisement prohibitions regarding modelled and actual results: No-Action Letter, Clover Capital Management, Inc. October 28, 1986

[2] See Backtesting, Harvey, Liu, 2015

[3] How to spot backtest overfitting? https://www.davidhbailey.com/dhbtalks/battle-quants.pdf , The Probability of Backtest Overfitting: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2840838 and Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample performance, Baley, Borwein, Prado, Zhu, 2013

[4] Predicting and preventing overfitting of financial models, Chalana, 05/2017 https://sites.math.washington.edu/~morrow/336_17/papers17/akshay.pdf

[5] See Backtesting, Harvey, Liu (07/2015), also see Appendix A

- Does the strategy assume that trades can be implemented at the same closing process as the one generating the trading signal, or is some delay/“slippage” accounted for?[6]
- What other assumptions are being used? (i.e. inclusion of transaction cost, fees, financing cost, stock lending fees, etc.)
- What transaction fees have been used in the backtest?

4: Backtesting time series

Assessing length of backtesting time series and cross-sectional approach

- Has the longest available dataset been used? If not, why not?

  Observation: there may be a lot of ways to define what the longest available dataset is (data for all markets available vs. some markets being available) and using the absolute longest may not always be the most representative approach.

- Have all the available markets in the asset class been tested? If not, why not?

  Observation: All the markets in an asset class should adhere to the risk premium (or at least not be counter indicative) irrespective of the liquidity level.

- Was a hypothesis based on an economic rationale that had been formed prior to backtesting? If not, why not?

  Observation: There should have been ranges of parameters that are reasonable that have been formed prior to backtesting.
- What is the sensitivity of the model to changing the parameters/markets/history[7]?
- Does the backtest use any proxy data? If so, what assumptions and adjustments have been made (and potential impact of such versus actual data)?

5: Backtesting results versus actual performance

- Disclosure of back-testing results (for launch, and subsequent strategy adjustments)[8] – see orange boxes in illustration below
- Disclosure of realised track record (between adjustment intervals)[9] – see blue boxes in illustration below
- Disclosure of “ghost” performance (for previous strategy implementations)[10] – see green boxes in illustration
- Are the environments when strategy performed well/badly in backtests similar to those while the strategy is “live”?

[6] An uncertainty quantification framework for the achievability of backtesting results of trading strategies Raymond Hon-Fu Chan, Alfred Ka-Chun Ma and Lanston Lane-Chun Yeung (https://www.risk.net/journal-of-investment-strategies/5331631/an-uncertainty-quantification-framework-for-the-achievability-of-backtesting-results-of-trading-strategies)

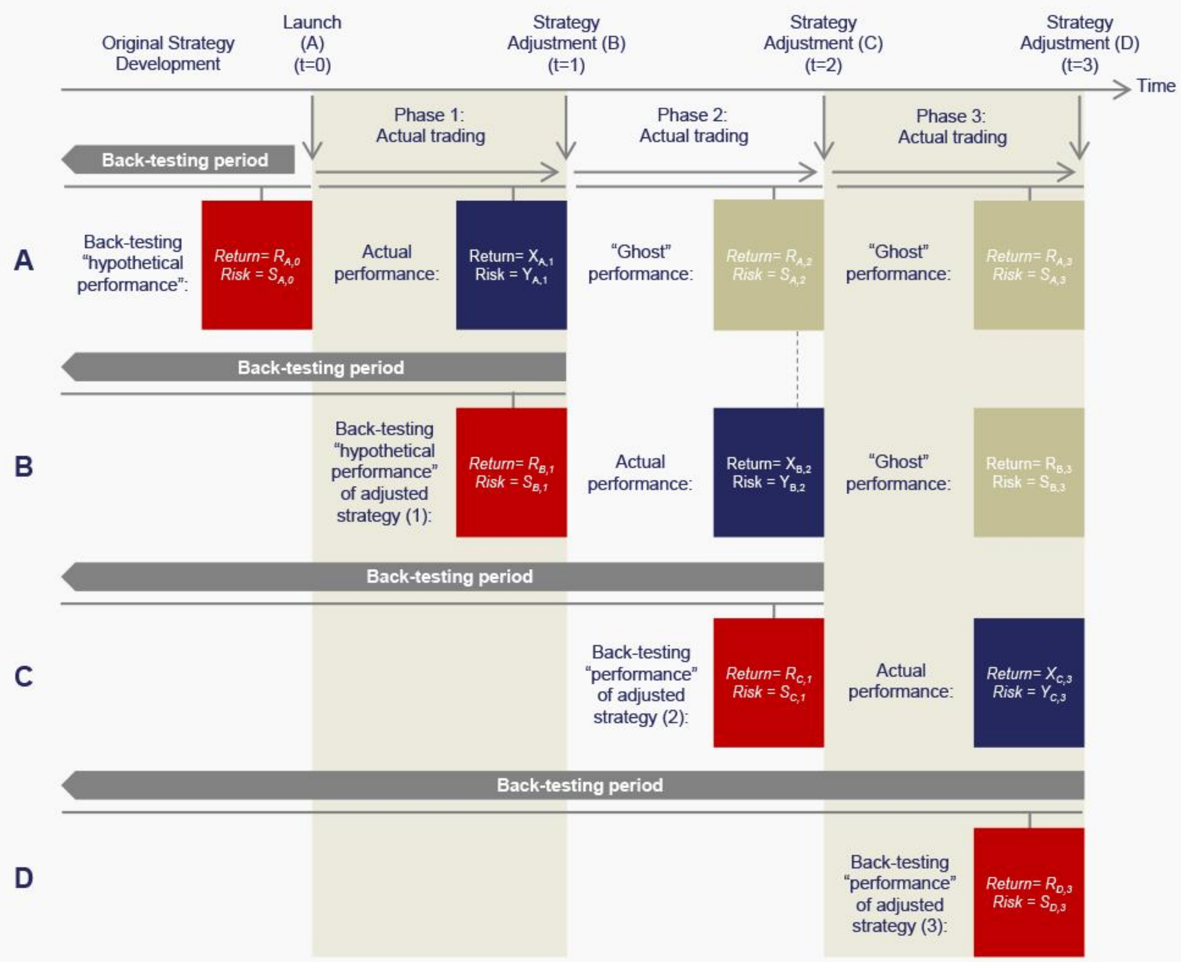
[7] E.g., a strategy looks good as a general average, but returns of the strategy occurred mainly in the distant past, but not more recently, or time varying market betas of factors.

[8] Where significant alterations in the strategy / overall objectives /design are being made, new back tests are required, but not for gradual (small) alterations

[9] Where a realised track record constitutes a carve out return (subcomponent of a fund), the carve out returns are not necessarily achievable on a stand-alone basis (e.g. different approach to risk management such as different draw down controls, different diversification benefit)

[10] Caveat: where certain input factors cease to be available, it might not be possible to keep a strategy running (e.g. LIBOR, discontinued markets or other input factors, etc.)

Illustration: Time series of relevant back-testing, actual performance and ghost performance results


Original Strategy Development
Launch (A) (t=0)
Strategy Adjustment (B) (t=1)
Strategy Adjustment (C) (t=2)
Strategy Adjustment (D) (t=3)
Time
Phase 1: Actual trading
Phase 2: Actual trading
Phase 3: Actual trading
Back-testing period
A
Back-testing "hypothetical performance":
Return= $R_{A,0}$ Risk = $S_{A,0}$
Actual performance:
Return= $X_{A,1}$ Risk = $Y_{A,1}$
"Ghost" performance:
Return= $R_{A,2}$ Risk = $S_{A,2}$
"Ghost" performance:
Return= $R_{A,3}$ Risk = $S_{A,3}$
Back-testing period
B
Back-testing "hypothetical performance" of adjusted strategy (1):
Return= $R_{B,1}$ Risk = $S_{B,1}$
Actual performance:
Return= $X_{B,2}$ Risk = $Y_{B,2}$
"Ghost" performance:
Return= $R_{B,3}$ Risk = $S_{B,3}$
Back-testing period
C
Back-testing "performance" of adjusted strategy (2):
Return= $R_{C,1}$ Risk = $S_{C,1}$
Actual performance:
Return= $X_{C,3}$ Risk = $Y_{C,3}$
Back-testing period
D
Back-testing "performance" of adjusted strategy (3):
Return= $R_{D,3}$ Risk = $S_{D,3}$

(A, B, C, D are the respective strategy implementations at times 0, 1, 2 and 3.)

Key questions for investors to ask

1. Disclosure

- Has the provider explained how the backtesting results were derived, that the result is not the performance of any actual account and that it is not a guarantee of future results?
- Has the track record been separated between simulated/theoretical performance and realized/live?
- Have all the strategy adjustments been correctly disclosed?
- If the strategy has been adjusted since launch, which strategy implementation has been used for showing the simulated track record (prior to launch) and why?
- How do the strategy implementations differ in history? (backtest, live and ghost performance)
- Is any of the historic realised performance based on carve out returns?

2. Back-testing process

Assumptions

1. Was a hypothesis formed prior to backtesting? If yes, what is it (and list academic references/rationale if relevant), if not why?
2. Does the strategy assume that trades can be implemented at the same closing price as the one generating the trading signal, or some delay/“slippage” accounted for?
3. What other assumptions are being used? (i.e. inclusion of transaction cost, fees, financing cost, stock lending fees, etc.)
4. How are the transaction costs incorporated in the backtest (e.g. which bid/ask spreads are used)?

Data

1. Have all the available markets in the asset class been tested? If not, why? What is the impact on the strategy by adding all of the markets?
2. Has the longest available dataset been used? If not, why? What is the performance of the strategy if it is back-extended?
3. What is the sensitivity of the performance by changing the parameters/markets/history?
4. Are any of the datasets proprietary or otherwise exclusive to the manager?
5. Are any of the datasets proprietary to third parties, such that they could become unavailable in the future?

Approach

1. Has the strategy been tested on out of sample data? (applying the strategy to data that has not been used in the initial backtesting phase)
2. What backtesting techniques have been employed to avoid over-fitting (train test split, multiple train text split, rolling window approach…)?
3. How many trials have been undertaken to come up with the strategy? (incl. looking at minimum backtest length for a given number of trials)
4. Have the metrics such as probability of backtest overfitting, performance degradation and probability of loss, stochastic dominance, etc been evaluated?
5. Has the performance (and risk) impact of strategy enhancements (“naked” versus enhanced strategy back-testing performance) been assessed?
6. How many degrees of freedom does the model contain[11] and what is the risk of overfitting due to excessive complexity or number of parameters?
7. Is the strategy expected to perform the same at scale? Are there capacity limitations? How was this evaluated?

[11] The model’s degrees of freedom correspond to the number of coefficients estimated minus 1

3. Interpretation

- What is an appropriate discount factor to use for the particular back-tested track record?
- What were the environments where the strategy performed well/badly in the backtests? (Same question for realized/live)
- If realised/live performance deviates from backtest results in similar market environments, what accounts for the difference?
- Is there anything significantly different in the current market conditions compared to the backtest which could have an impact on the strategy going forward?
- What is the back-tested track record of the model when the investor provides any set of parameter values?

Appendix A

Other areas of asset management / finance where back-testing is being used

Back-testing is being used in many areas of finance. Existing industry practices, regulatory guidance and academic literature provide insights about how good back-testing practices should look.

Key areas of focus

- Prevention of deceptive communication (e.g. mixing back-testing results with actual performance), requirement to clearly label back-tests
- Approaches to discounting back-test Sharpe Ratios
- Prevention of false assumptions regarding “tradeable prices” (i.e. using a closing price as a trading signal and simultaneously tradeable prices)
- Assessing / understanding dispersion of indices which seek to model similar underlying risk premia
- Ongoing comparison of model projections against realised values (applicable in context of banking risk models)

See below for overview of regulations and academic papers.

Regulations/practices

Source	Content/approach
US: Advisers Act Rule 206(4)-1	Prohibition of fraudulent, deceptive, or manipulative (communication) practices
No-Action Letter, Clover Capital Management, Inc. October 28, 1986	Guidelines for advertising with actual and model performance. The letter specifically lays out the standards the SEC staff uses to determine whether the advertising is fair and not misleading. Prohibits mixing models/back-tests with actual performance.
CFA Institute: GIPS Standards	• Hypothetical and back-tested composite returns do not satisfy the requirements of the GIPS standards • To be GIPS compliant, performance data must only contain actual portfolios managed by the firm • Hypothetical or back-tested results can only be included when clearly labelled as supplemental information
Basel II: Sound practices for backtesting counterparty credit risk models	Focus on the quantitative comparison of the IMM[12] banks models’ forecasts against realised values.

Select research papers

Source	Content/approach
Alternative Risk Premia: Is the Selection Process Important? The Journal of Wealth Management, 22 (1) 25-38 (Summer 2019). Francesc Naya, Nils Tuchschmid	• Many ARP indices have been proposed by different providers that claim to capture the same underlying risk premia. Some of these categories of indices show risk-return characteristics that are rather homogeneous, others are highly heterogeneous. Hence, performance is provider dependent making the choice of an index an important component of the allocation process • A proposed index may not automatically mimic an existing risk premium whose performance is sustainable or persistent: Differences between simulated past results and live data for individual indices suggest significant overfitting bias. Once

[12] Internal Model Method

	launched, the performance of ARP indices dropped significantly - Conclusion: When it comes to allocating capital to ARP, an extensive due diligence/selection process is required
Backtesting, The Journal of Portfolio Management, 42 (1) 13-28 (Fall 2015) Campbell R. Harvey, Yan Liu	- Paper develops an analytical way to determine the magnitude of the haircut to be applied to back-test results (Sharpe Ratios) - It suggests that the "common" practice of discounting reported Sharpe Ratios of trading strategies by 50% (rule of thumb) is not adequate and should be replaced by a non-linear approach that only moderately penalises the highest Sharpe Ratios while the marginal Sharpe Ratios are heavily penalised
An uncertainty quantification framework for the achievability of backtesting results of trading strategies (Raymond Hon-Fu Chan, Alfred Ka-Chun Ma and Lanston Lane-Chun Yeung, (September 2012)	- Back-testing has always been indispensable in analysing the profitability of trading strategies in the empirical finance literature. When measuring return, while most of the literature implicitly assumes that a trade can be implemented at the same closing price as the one generating the trading signal, some empirical evidence has been found suggesting that this assumption presents a significant challenge to the robustness of their results - The results show that a significant number of technical trading strategies with positive returns are found to be unviable in the presence of implementation uncertainty
Quantifying Backtest Overfitting in Alternative Beta Strategies Journal of Portfolio Management Vol. 43, Nr. 2 (Winter 2017) Antti Suhonen (Aalto University School of Business), Matthias Lennkh (Clear Alpha Limited), Fabrice Perez (Clear Alpha Limited)	- Assessment of the biases in the back-tested performance of "alternative beta" strategies using a sample of 215 commercially promoted trading strategies across five asset classes - Results lend support to the cautions in recent literature regarding back-test overfitting and lack of robustness in trading strategy performance during the "live" period (out of sample) - Median 73% deterioration in Sharpe ratios between back-tested and live performance periods for the strategies in our sample - Establishment of a link between performance deterioration and strategy complexity, with the realized reduction in live vs. back-tested Sharpe ratios of the most complex strategies exceeding those of the simplest ones by over 30 percentage points - Robustness of strategy exposure to risk factors varies between asset classes and strategies, and appears reasonable in equity volatility and FX carry strategies, but quite weak in the equity value strategy in particular
Alice's Adventures in Factorland: Three Blunders That Plague Factor Investing (Arnott, Harvey, Kalesnik, Linnainmaa)	- Paper assesses problems that might be underappreciated by investors (factor performance expectations, downside risks, diversification)

Appendix B
Working group members

Name	Title	Organisations
Iivo Paukkeri	Portfolio Manager	Aalto University Foundation
Duncan Moir	Senior Investment Manager, Alternative Investment Strategies	Aberdeen Asset Managers Limited
Avgustina Sarkizova Evelina Klerides	Partner, Dynamic Beta Partner, Dynamic Beta	Albourne Partners
Walter Cegarra	Founder	Arch Ventures
Deepak Gurnani	Founder	ARP Americas
Christopher Reeve	Director of Risk	Aspect
Andre Breedt	Research Associate	Capital Fund Management
Apostolos Katsaris	CIO	CdR Capital Ltd
Melissa Hill	Co-Founder	Eleos Capital Advisors Limited
Nicolas Papageorgiou	CIO, Public Markets	Fiera Capital
Hugues Bessette	Chief Investment & Risk Officer	Innocap
Steven Desmyter	Global Co Head Sales & Marketing, Man Group and Global Co Head of Responsible Investing	Man Group
Lisa Fridman	Portfolio Manager	Martlet Asset Management
Scott Treloar	CEO	Noviscient
Matt Talbert	Senior Investment Manager	Teacher Retirement System of Texas
Jerome Teiletche	Head of Cross Asset Solutions, Managing Director	Unigestion
Samantha Foster	Managing Director, Investments Office	USC University of Southern California
Dr. Sushil Wadhwani	CIO	QMA Wadhwani
Neal Howe	Partner & Director of Investor Solutions	Welton Investment Partners
Rodney Livingston	Senior Investment Officer	West Virginia Investment Management Board
Thomas Deinet	Executive Director	SBAI