Algorithmic Trading (COMP0051)

Paolo Barucca Nick Firoozye

Department of Computer Science
University College London

12 February 2020

Algorithmic Trading (COMP0051)

- email: p.barucca@ucl.ac.uk - n.firoozye@ucl.ac.uk
- office: 66-72 Gower Street, room 4.08 (fourth floor)
- office hour: Friday 11am

Data snooping and backtest overfitting

With four parameters I can fit an elephant, and with five I can make him wiggle his trunk.

-Attributed to *John von Neumann* by Enrico Fermi,
as quoted by Freeman Dyson in *A meeting with Enrico Fermi* in Nature 427 (22 January 2004) p. 297

Backtest overfitting = P-hacking = Data Snooping

- P-hacking is a problem in science
- Lack of reproducible results
- OOS performance completely different from IS performance

Standard methods:Test-Train-Holdout

- Standard ML/Statistical method is to divide into
  - Train data
  - Test data
  - Possibly rotating as in CV-based methods
  - Holdout data (not seen until the end)
- Cross-validation to determine appropriate model selection
  - *Seems* like using OOS data, no?
  - Is it?
- At least we have the holdout!
  - Holdout overfitting is a problem
  - Kaggle Competition Hacking
  - Google/IBM/Samsung/MS paper

Overfitting

Simpler is better?

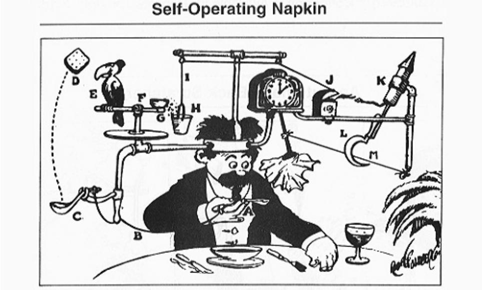


Figure: Rube Goldberg

Algorithmic Strategies

- Have their own *crosses to bear*
  - No model
  - No forecast
  - No likelihood
  - No standard errors

- Weights
  - Estimated directly
  - Not indirect, Forecast to Weight
  - May be more efficient than forecasting
- Judged by Sharpe Ratio / Calmar / Sortino
- Each of these, possibly optimised (IS)
- May lead to spurious results
- Poor OOS performance

Philosophical Guidance - Occam's Razor

Figure: Occam's Razor

Or as they say, "The principle states that among competing hypotheses that predict equally well, the one with the fewest assumptions should be selected."
Guardian Science

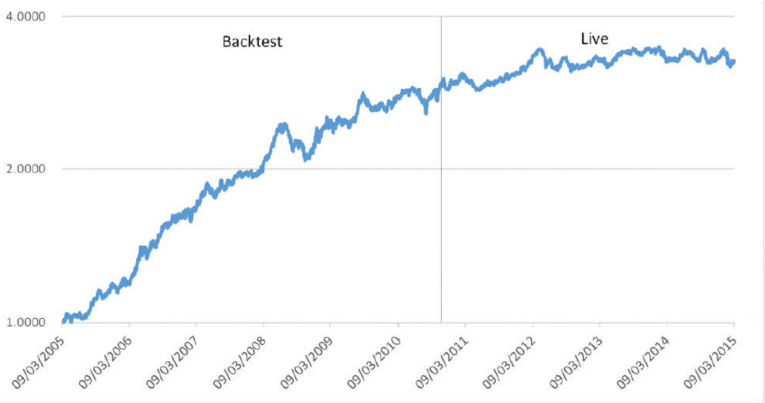


Figure: Clear Alpha Backtest

Suhonen et al, Quantifying Backtest Overfitting in Alternative Beta Strategies

Serious problem 2

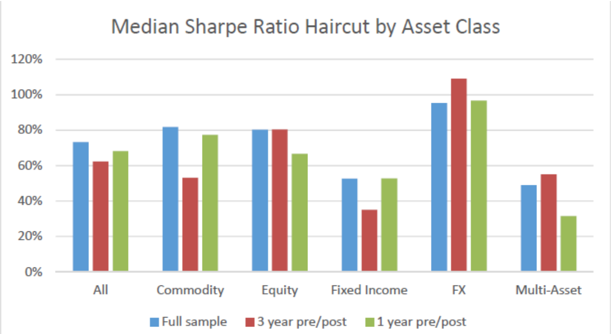


Figure: Clear Alpha Sharpe Ratio Haircut

Suhonen et al, Quantifying Backtest Overfitting in Alternative Beta Strategies

Financial Charlatanism - Bailey et al

- Show that if one chooses between N strategies of zero Sharpe ratio
- If sample size is small enough
  - Can find (high) non-zero Sharpe
- Try to determine minimum backtest length to ensure that for a given number of trials N
  - If the DGP has Sharpe of 0
  - The expected maximum Sharpe Ratio over a set of N strategies is bounded at 1
- Min BTL is more of an illustrative concept
- Awkward concept of ‘independent’ trials

Min back-test length

- Number of trials allowed for given back-test length
- To prevent maximum Sharpe of skill-less strategies to be over 1

Pseudo-Mathematics and Financial Charlatanism

Adjusted Sharpe-Ratios - Harvey-Liu

- Multiple hypothesis testing framework
- No need for *independence* of hypotheses (whatever that means)
- Need to know
  - Sample size T
  - Number of models under consideration N
- Sharpe Ratio = t-statistic/$\sqrt{T}$ can be converted into p-value
- Adjust p-values to recomputed *Adjusted Sharpe Ratios*

Sharpe Ratio Statistics

- Andrew Lo, The Statistics of Sharpe Ratios gives large sample statistics
- In small sample a single Sharpe has a Student *t* distribution

  $$p^I = Pr(r > SR\sqrt{T})$$

  is given by the t_{T-1} distribution (i.e, with $T - 1$ df).
- If underlying DGP has *true* Sharpe of zero - what does a SR of 5% mean?
- Example: $p = 0.05$ Want to accept at 5% confidence
  - If $M = 1$ (one independent test) - false positives will occur only 5% of the time
  - If $M = 10$, then $p^M = 1 - (1 - p)^{10} = 1 - 0.95^{10} = 0.401$ and at least one false positives will occur close to 40% of the time
  - Since they are not independent, we need to adjust CI
- Adjust Sharpe so they retain significance and we do not accept more false-positives

Multiple tests in Statistics

- Assuming there are M hypotheses
- We have generated $(p_1, \ldots, p_M)$
- If we reject (R) hypotheses, we have
  - let N_r be the false discoveries (strategies incorrectly classified as profitable)
- Multiple approaches to reducing false-negatives

Multiple approaches to dealing with multiple tests

- Two different types of false-negatives:
  - **Family Wise Error Rate (FWER)**

$$FWER = Pr(N_r \geqslant 1)$$

    - Wish to reduce the chance of exactly one or more errors.
    - This means almost no loss-making strategies
    - Makes sense in physical sciences or space-missions.
  - **False Discovery Proportion rate (FDP)** $FDP = N_r/R$ if $R > 0$ and 0 otherwise.
  - **False Discovery Rate (FDR)** is $FDR = E[FDP]$
    - Wish to keep proportion of false discoveries the same irrespective of total number of tests
    - allows number of bad strategies to grow but proportion to remain the same
    - More appropriate to our risk-taking approach

- An adjustment for *FWER*

$$p_i^B = \min[Mp_i, 1], ; i = 1, \ldots, M$$

- Example: 10% CI, three tests
  - In ascending order, $P = (0.02, 0.05, 0.20)$,
  - Accept the first two?
  - Bonferroni adjustments
    $P^B = (3 * 0.02, 3 * 0.05, 3 * 0.20) = (0.06, 0.15, 0.60)$
  - Only accept the first p-value amongst the three.

- Another adjustment for *FWER*:

$$p_i^H = \min[\max_{j \leqslant i}[(M - j + 1)p_j], 1], \quad , i = 1, \ldots, M$$

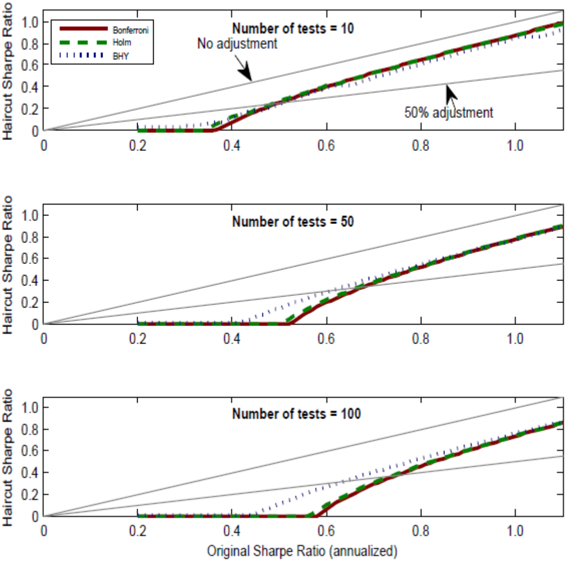
- Sequential testing method
- Less arduous than Bonferroni.
- Example: same three p values as before, we would get
  - Holm adjustments
    $P^H = (3 * 0.02, 2 * 0.05, 0.20) = (0.06, 0.10, 0.20)$.
  - Accept both the first and the second p-values as being significant at the 10% level.
- We note that $p_{(i)}^{Holm} \leqslant p_{(i)}^{Bonferroni}$ for all i
- Both Holm and Bonferroni are made to ensure that we reduce FWER so that it is below our original confidence level.

- Benjamini-Hochberg-Yekutieli (BHY) meant to reduce *FDR* with the following recursively (from highest p-values to lowest) defined weights:

$$p_i^{BHY} = \begin{cases} p_{(M)} & if \text{i= M} \\ \min[p_{i+1}^{BHY}, \frac{M \times c(M)}{i} p_i] & if \text{i} \leqslant M-1 \end{cases}$$

  where $c(M) = \sum_{j=1}^{M} \frac{1}{j}$
- Example: $P = (0.02, 0.05, 0.20)$ is ordered p-value sequence
- $c(3) = (1 + \frac{1}{2} + \frac{1}{3}) = \frac{11}{6}$
- $P^{BHY} = (0.11, 0.14, 0.20)$ accepts none as significant at 10%
- Somewhat atypical, since BHY is often less arduous than Holm and Bonferroni
- For larger number of tests, $p^{BHY} \leqslant p^{Holm} \leqslant p^{Bonferroni}$
- Works with arbitrary dependence structure of p values

Adjusted Sharpe Ratios



What can we do in Practice?

- **Strong Priors** Only found via domain specific knowledge and experience. Intuition
- **Panel Data** (different data, technically independent test results)
  - Different
    - Currencies
    - Exchanges
    - Tenors
    - Expiries
    - Futures, etc
  - Many reasons for differences-institutional circumstances, etc
  - Explain differences
  - Isolate what you're being compensated for
- **Bootstrap** Randomized tests of continued performance (e.g., Halbert White, A Reality Check for Data Snooping) is far too arduous–tests continued outperformance (e.g., *stochastic dominance*)

Extra Readings

- Bailey et al, Pseudo Mathematics and Financial Charlatanism Interesting but inconclusive.
- Harvey-Liu, Backtesting Much more sound statistically, but much more humble title
- Harvey-Liu, Evaluating Trading Strategies Layman's version of Backtesting
- Quantifying Backtest Overfitting in Alternative Beta Strategies Displays many bank *smart-beta* strategies, IS vs OOS.

P-Hacking in Science

- Gelman - Blog post on Garden of Forking Paths
- Gelman-Loken on Garden of Forking paths
- Other refs FAQs on Backtest Overfitting

Reusable Holdout

- Dwork et al, Reusable Holdout (2016) - Differential Privacy - a carefully constructed randomized sample to ensure non-invertible tests. This is closely related to work by Blum-Hardt on Kaggle-competition hacking.
- New paper by Google, MS, IBM, Samsung, et al authors
- Blog-post 1
- Blog post 2
- Lecture Notes
- Dwork et al, Preserving Statistical Validity in Adaptive Data Analysis
- Dwork et al, Generalization in Adaptive Data Analysis and Holdout Reuse

Data-snooping

- Halbert White – Data Snooping tests (use stationary block bootstrap - random blocksize, and a given test statistic to look at likelihood of bootstrap). This was one of the first studies (2000). It is more like a test of stochastic dominance than what is truly appropriate for our considerations
- Sullivan-Timmerman-White, Data Snooping Reality Check Overly arduous bootstrap test of trade efficacy (using FWER). Almost every trade fails this. Is it an appropriate objective?
- Github Repository - Data Snooping tests Quite a few tests with relevant papers, coded in matlab