Detection of False Investment Strategies through FWER and FDR

Prof. Marcos López de Prado

Advances in Financial Machine Learning

ORIE 5256

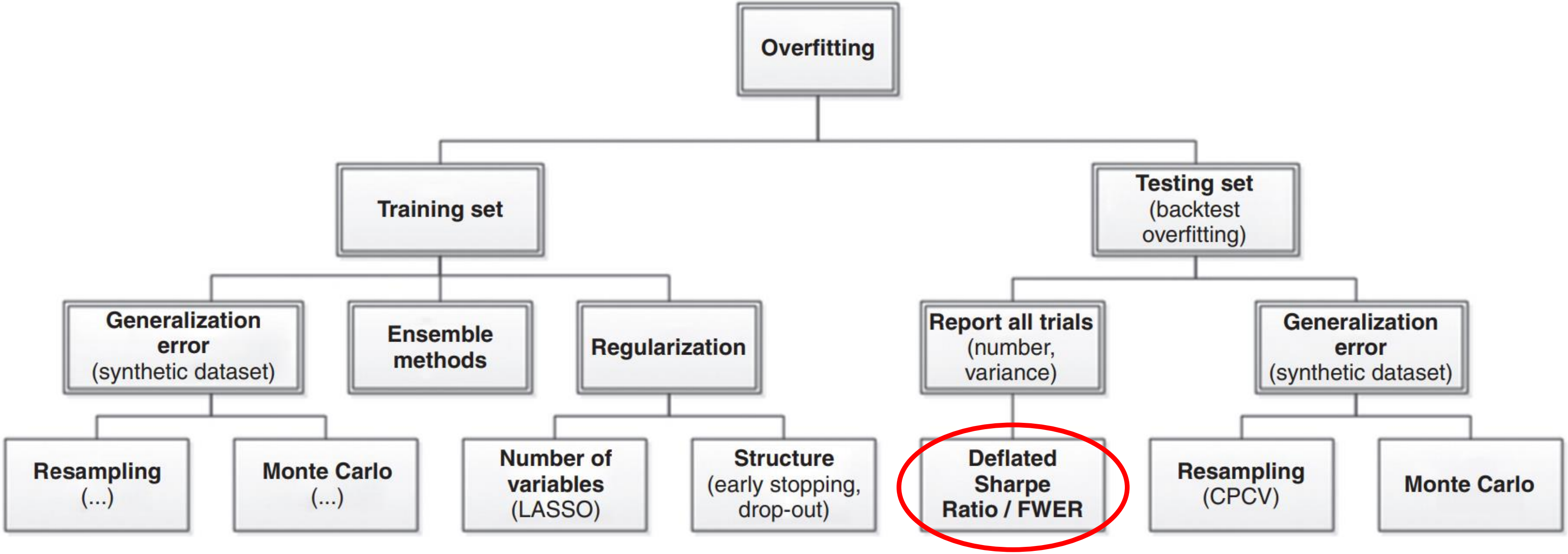
Preamble (1/3)

- Financial systems rarely allow experimentation
  - E.g., we cannot reproduce the flash crash of 2010 while controlling for environmental conditions
- As a result, much financial research relies on the statistical analysis of finite (historical) datasets, where:
  - Time series datasets are limited
    - It takes 10 years to produce 10 years of history
    - Observations are serially-conditioned
  - The investment universe is limited
    - The number of securities is often in the tens of thousands, however their cross-section is conditioned by very few common factors
- Finite datasets may be overfit in two ways:
  - Train-set overfitting. Solutions include:
    - Estimation of the generalization error, e.g. resampling (cross-validation) methods, Monte Carlo, …
    - Regularization, e.g. Tikhonov (Ridge) regression, LASSO, elastic nets, early stopping, drop-out, …
    - Ensemble methods, e.g. bagging, boosting, stacking, …
  - Test-set overfitting

Preamble (2/3)

- Test-set overfitting is particularly difficult to avoid
  - One typical example in finance is backtest overfitting
- Backtest overfitting occurs when a researcher backtests multiple investment algorithms, and selects the best performing one
  - This kind of overfitting occurs even when the backtests are accurate representations of historical performance, using point-in-time information
- Solutions to test-set overfitting include:
  - Controlling for FWER and FDR targets, e.g. Sidak, Holm, Benjamini-Hochberg
  - Deflating the *p*-values, e.g. Deflated Sharpe ratio (DSR), Romano-Wolf *p*-value correction
  - Estimating the generalization error, e.g. resampling (CPCV) backtesting, Monte Carlo backtesting
- **This presentation focuses on controlling for FWER and FDR in the context of selecting among multiple investment strategies**

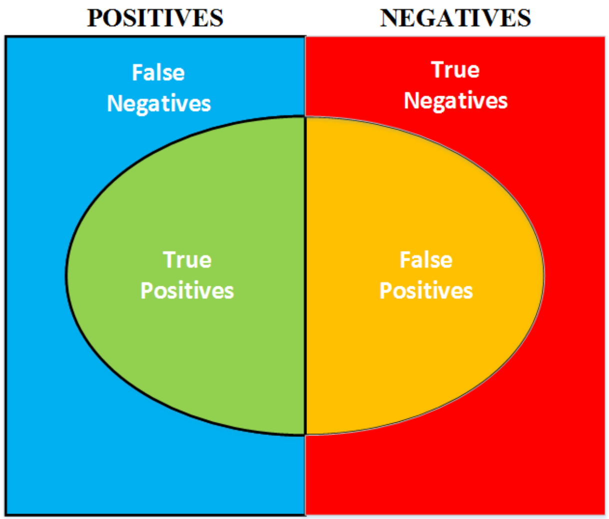
Preamble (3/3)


Overfitting
Training set
Testing set (backtest overfitting)
Generalization error (synthetic dataset)
Ensemble methods
Regularization
Report all trials (number, variance)
Generalization error (synthetic dataset)
Resampling (...)
Monte Carlo (...)
Number of variables (LASSO)
Structure (early stopping, drop-out)
Deflated Sharpe Ratio / FWER
Resampling (CPCV)
Monte Carlo

Statistical Testing of Hypotheses

Confusion Matrix

- In a binary classification problem, we wish to assign labels to observations
  - **Negative** labels are typically associated with the normal state of nature
  - **Positive** labels are typically associated with abnormal observations, which require action
- A binary classifier is an algorithm that predicts which observations are positive (and by default, which are negative)
- Combining the prediction with the ground truth, gives:
  - True Positive (TP)
  - False Positive (FP): Type I Error
  - True Negative (TN)
  - False Negative (FN): Type II Error

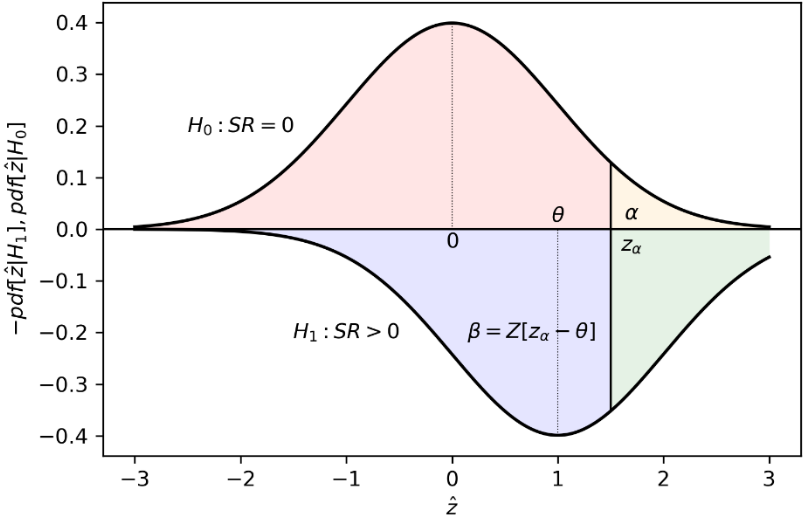


A visual representation of the confusion matrix:

- False positive rate (FPR): FP/(FP+TN)
- Confidence: 1-FPR
- False discovery rate (FDR): FP/(FP+TP)
- Precision: 1-FDR
- Recall, Power: TP/(TP+FN)

The Neyman-Pearson Framework [1933]

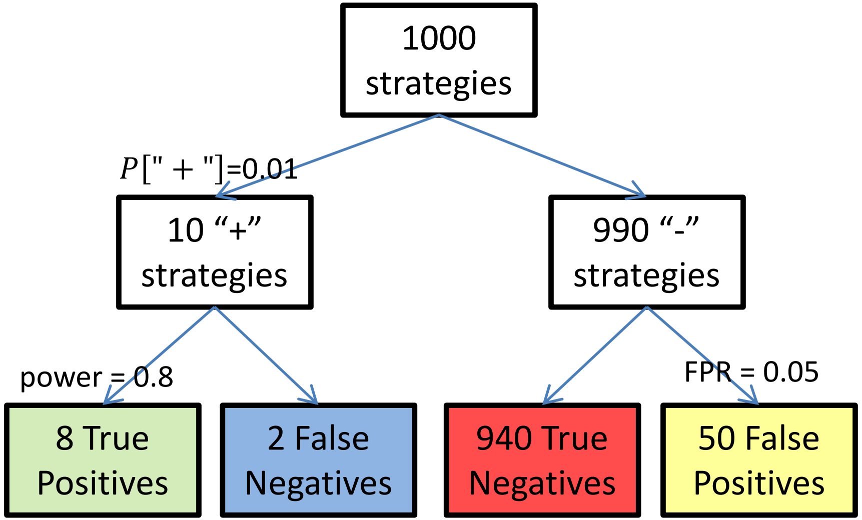
- Jerzy Neyman and Egon Pearson proposed a framework for testing which of two competing hypotheses was supported by empirical evidence
  - The test's statistic for **Negative** cases follows a known distribution characterized by the null-hypothesis (H_0)
  - The test's statistic for **Positive** cases follows a known distribution characterized by the alternative hypothesis (H_1)
- Powerful tests separate the two distributions
- Under various assumptions, this implies a confusion matrix that controls for either
  - FPR (set by α), or
  - Power (set by β)



Hypothesis testing aims at rejecting H_0 given an evidence z, while controlling for FPR (and implicitly for Power). The above plot is colored like the Confusion Matrix. FPR is the ratio between the yellow area and the area above zero, whereas FDR is the ratio between the yellow area and the area to the right of z_α. Hence, computing FDR requires knowledge of negatives as well as positives.

Caveats of Classical Hypothesis Testing

FPR vs FDR



1000 strategies

$P[" + "]$=0.01

10 “+” strategies

990 “-” strategies

power = 0.8

FPR = 0.05

8 True Positives

2 False Negatives

940 True Negatives

50 False Positives

Suppose that the probability of a backtested strategy being profitable is 1%.

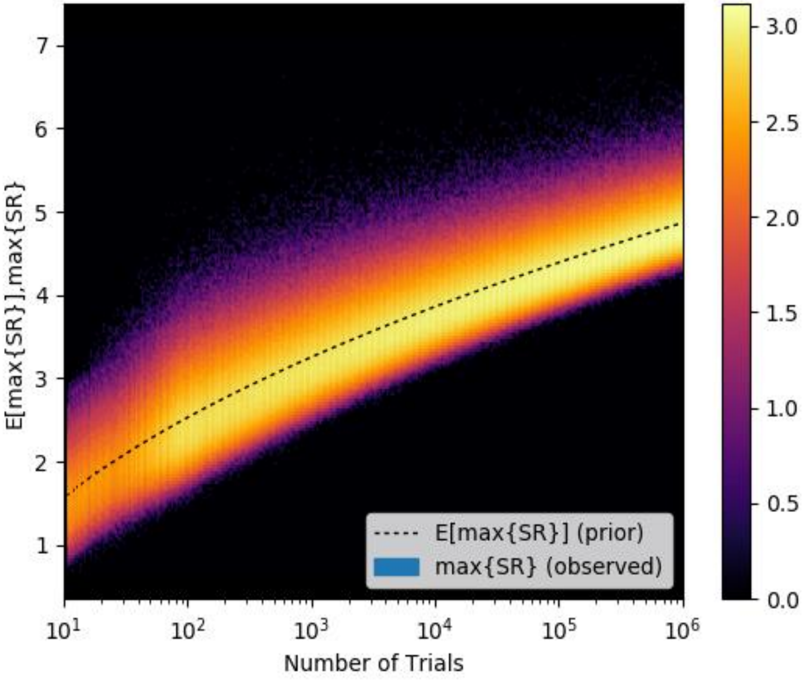
Then, at the standard thresholds of 5% significance and 80% power, researchers are expected to make 58 discoveries out of 1000 trials, where 8 are true positives and 50 are false positives.

Under these circumstances, **a p-value of 5% implies that at least 86% of the discoveries are false!** (…even if we do not select one out of many)

The above example illustrates why controlling for FPR can be misleading in finance, when positive conditions are a relatively rare event. In general, researchers should control for FDR ($P[H_0|z > z_\alpha]$) rather than FPR (or *p*-values, $P[z > z_\alpha|H_0]$). However, when researchers select the strongest positive out of many, FWER is more appropriate.

Multiple Testing

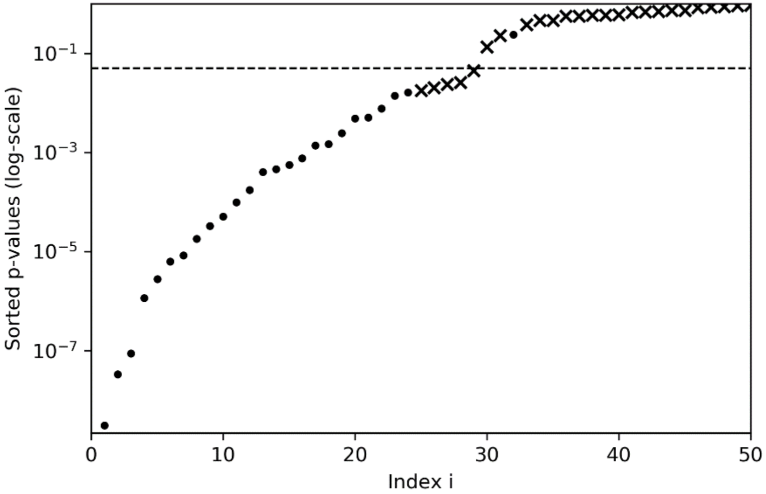
- Selection bias occurs when a researcher
  1. tests multiple hypotheses (trials)
  2. selects one or several positives
  3. presents that result as if other tests had not taken place
- There are two fundamental solutions to multiple testing
  - Control for FWER (or FDR, if several trials selected)
    - E.g., Bonferroni, Sidak, Holm, Hochberg, …
  - Adjust *p*-values
    - E.g., Romano-Wolf, Bailey-Lopez de Prado, …
- In this presentation, we focus on FWER / FDR
  - For a discussion of *p*-value adjustment in the context of strategy selection, see here



Distribution of the maximum estimated Sharpe ratio as a function of the number of trials, where the true Sharpe ratio is zero. See “The False Strategy Theorem”.

A Numerical Example

- Consider a set of 50 uncorrelated strategies, where we expect that
  - half of them have an annualized Sharpe ratio of 1
  - half of them have an annualized Sharpe ratio of 0
- Under the assumption of iid Normal returns, we estimate *p*-values on their mean returns
  - Positive conditions are marked with a circle
  - Negative conditions are marked with a cross
  - The dashed line shows the unadjusted rejection threshold that targets a FPR of 5% (α=0.05)
- <u>Question</u>: Can we separate positive strategies from negative strategies?



Selection bias occurs when we choose one or more statistical results out of many, and discard the rest (as if the discarded tests had not taken place). In the example above, at $\alpha = 0.05$, the test rejects 30 null hypotheses, with 5 false positives and 1 false negative ($FDR = 5/30 \approx 17\%$). As we increase the number of tests, FDR will rise, even if we keep FPR constant.

Controls for FWER

What is FWER?

- We denote each sample $\{x_i\}_{i=1,\dots,I}$ as a trial
- Consider K such trials, $\{x_{i,k}\}$, where $i = 1, \dots, I$ is the index of observations for trial k, and $k = 1, \dots, K$
- Let S_0 be the set of indices of true null hypotheses, having K_0 members
- For each trial k, we can estimate a *p*-value p_k
- The Familywise Error Rate (FWER) is defined as the probability of obtaining at least one false positive

$$FWER = Pr\left[\bigcup_{k \in S_0} (p_k \leq \alpha)\right]$$

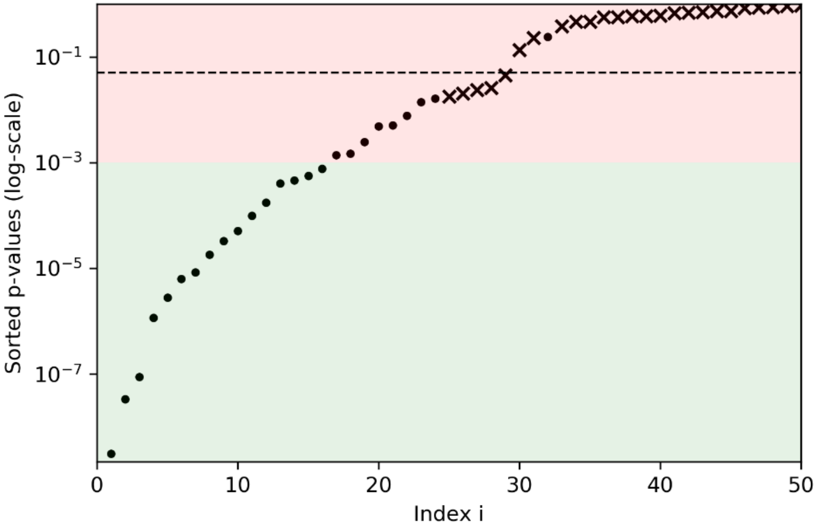
Bonferroni [1936]

- This method rejects all null hypotheses for which their respective *p*-values fall below $\frac{\alpha}{K}$
- That boundary controls FWER at level α, because

$$FWER = Pr\left[\bigcup_{k \in S_0} (p_k \leq \alpha)\right]$$

$$\leq \sum_{k \in S_0} Pr\left[p_k \leq \frac{\alpha}{K}\right] = K_0 \frac{\alpha}{K} \leq \alpha$$

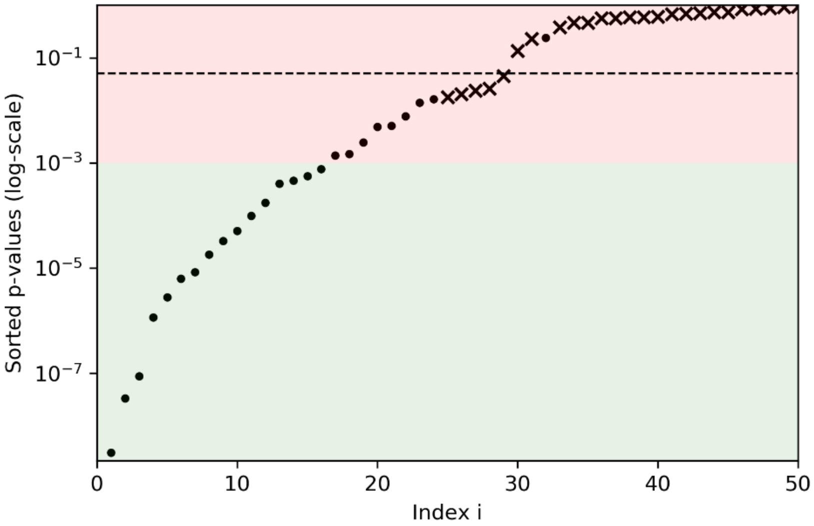
- The above statement relies on Boole's inequality, and it does not require the independence of the K trials



Applying Bonferroni's adjustment, the green area holds 16 rejected null-hypotheses, targeting a $FWER \leq 5\%$. This results in 0 false positives, at the cost of 9 false negatives.

Sidak [1967]

- This method tests each hypothesis at level $\alpha_K = 1 - (1-\alpha)^{1/K}$, instead of $\frac{\alpha}{K}$
- Under the assumption of independence, the procedure is more powerful than Bonferroni's, however the gain is small
- If the tests are negatively dependent, Sidak's correction can fail to control for FWER



Although Sidak's correction is more powerful than Bonferroni's, the gain is almost unnoticeable. Both methods are extremely conservative in their control of FWER.

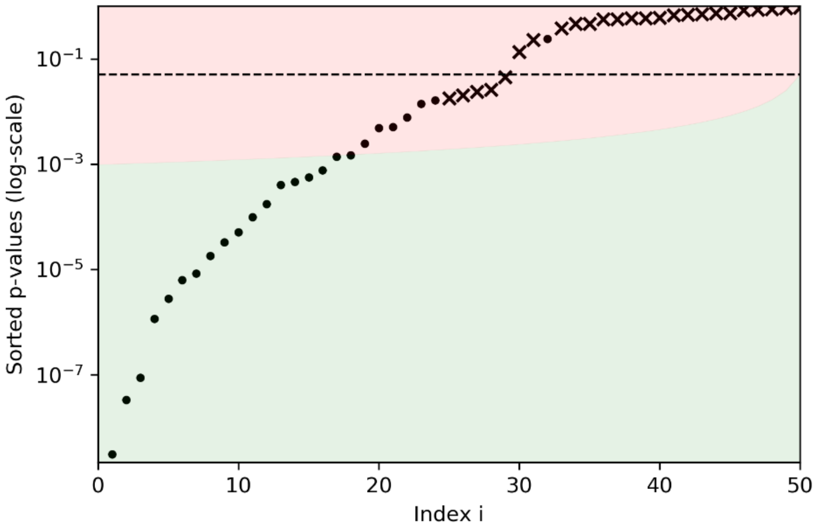
Holm [1979]

1. Sort all *p*-values in ascending order
2. For index k, find the smallest value k^* such that

$$p_k > \frac{\alpha}{K - k + 1}$$

3. Reject all null hypotheses associated with *p*-values $k < k^*$

- This method rejects all hypothesis rejected by Bonferroni, and potentially a few more
- Holm's method is uniformly more powerful than Bonferroni's (it produces fewer false negatives)



In our example, Holm's method produces 0 false positives and 7 false negatives (2 fewer than Bonferroni and Sidak).

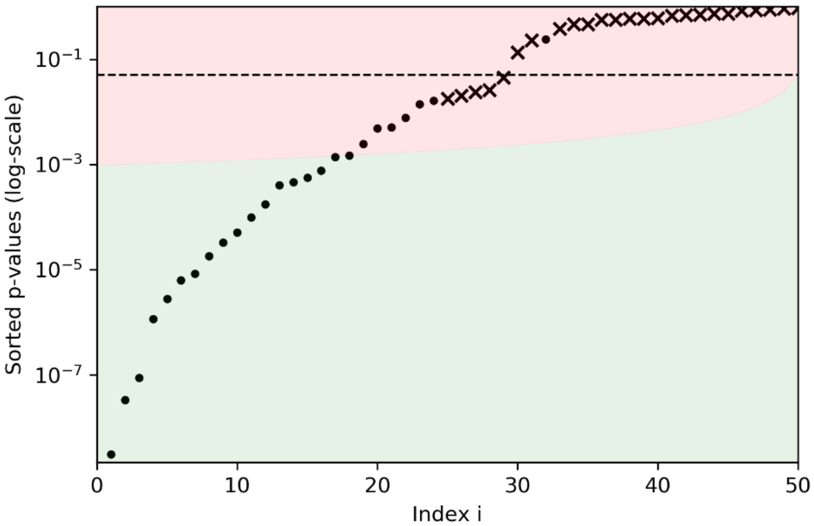
Hochberg [1988]

1. Sort all *p*-values in ascending order
2. For index k, find the largest value k^* such that

$$p_k \leq \frac{\alpha}{K-k+1}$$

3. Reject all null hypotheses associated with *p*-values $k \leq k^*$

- Hochberg's correction is more powerful than Holm's, however
  - it only holds under non-negative dependence
  - it can fail to control FWER if the tests are negatively dependent



In our example, Hochberg's correction leads to the same result as Holm's. The result would have been different if one or more *p*-values after index 18 fell within the green area.

Romano-Wolf [2005]

- The previous methods control for FWER by making various assumptions on the dependence structure across the *p*-values
- Those assumptions simplify the methods, at the cost of potentially lower power (a larger number of false negatives). For example:
  - the tests can be too conservative under positive dependence
  - Hochberg's test can fail to control for FWER under negative dependence
- To address this concern, Romano and Wolf proposed a method to control for FWER that takes into account the dependence structure of the test statistics, by resampling from the original data
- The procedure is computationally intensive
  - For an implementation, see Clarke, Romano and Wolf [2019]

Controls for FDR

Benjamini-Hochberg [1995]

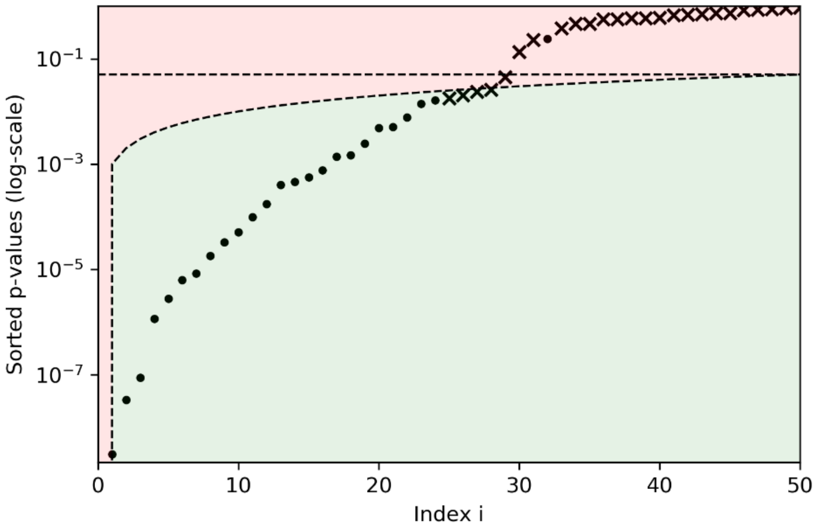
1. Sort all *p*-values in ascending order

2. For a target FDR q, find the largest index k^*

$$p_k \leq \frac{k}{K} q$$

3. Reject all null hypotheses associated with *p*-values $k \leq k^*$

- Because the method assumes independence of the tests,
  - it can fail to control for FDR when the tests exhibit negative dependence
  - it can be too conservative when the tests exhibit positive dependence



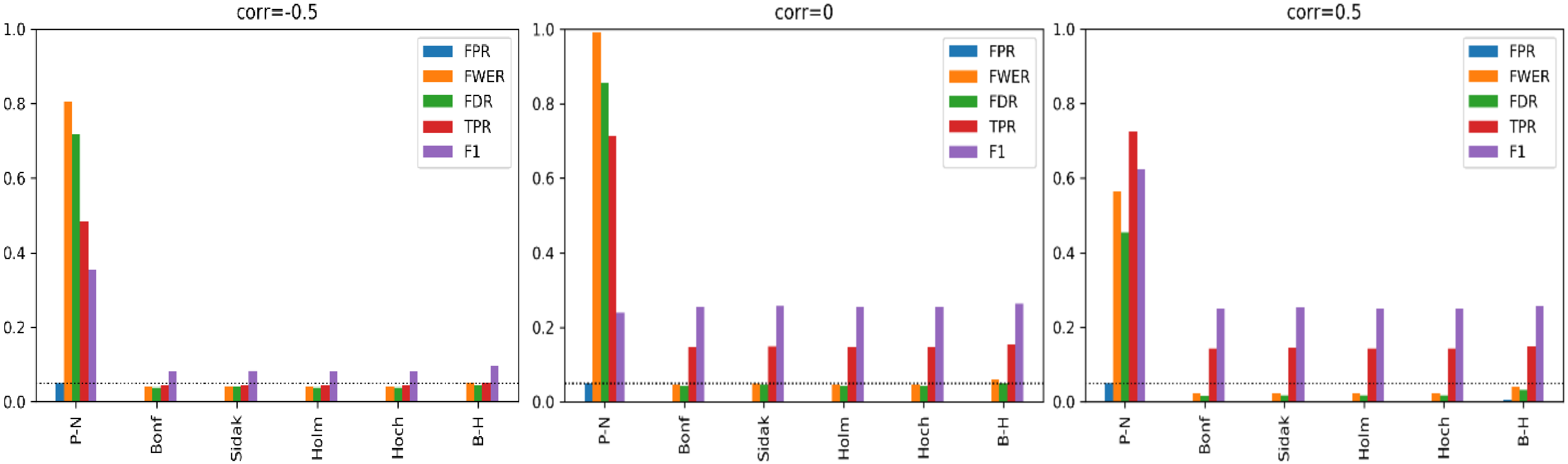
While targeting an FDR of 5%, Benjamini-Hochberg's correction rejects 29 null-hypotheses, resulting in 4 false positives and 1 false negative. As we can see, the FDR correction is less conservative than FWER, at the cost of accepting a greater number of false positives.

Experimental Results

Experiment Design

- Draw 10,000 simulations with
  - Number of trials: 100
  - Number of positive conditions: 1 (rare events)
  - Number of observations: 250, 1250, 2500 (equivalent to 1, 5, 10 years of daily observations)
  - Sharpe ratio scenarios: $0.5/\sqrt{250}$, $1/\sqrt{250}$, $2/\sqrt{250}$ (non-annualized)
  - Average correlation between trials: -0.5, 0, 0.5
- For each simulation, we generate the data for the 100 trials, estimate the *p*-values, apply the tests, classify results (TP, TN, FP, FN), compute scores (FPR, FWER, FDR, TPR, F1), and average scores across simulations
- This experiment allows us to compare interactions between 27 different scenarios

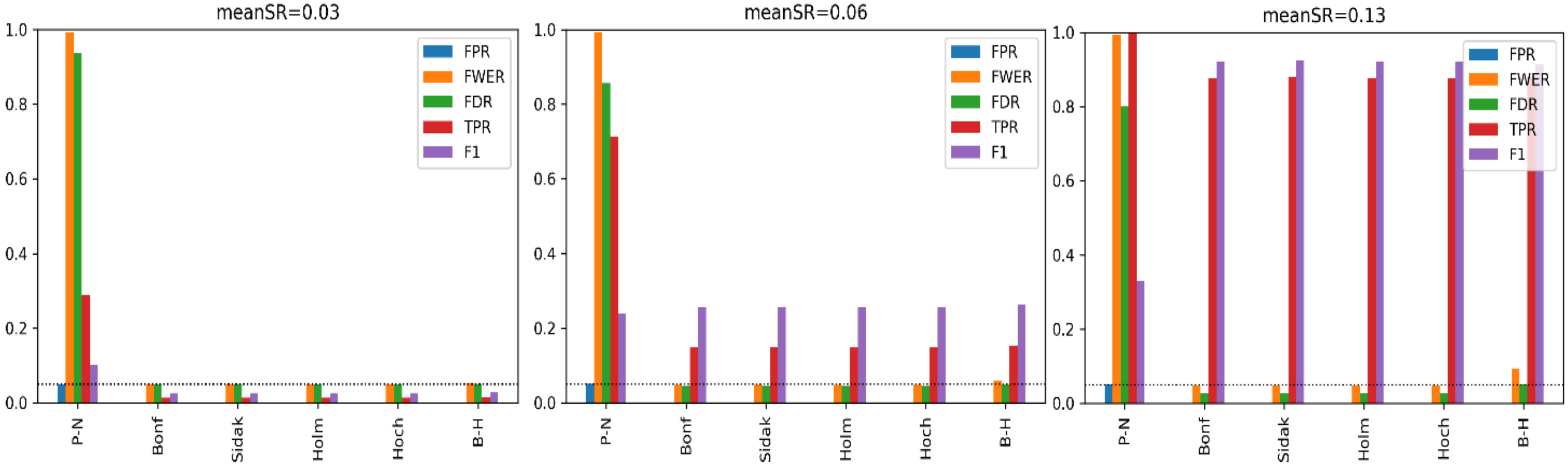
Correlation Effects



As expected, FWER and FDR are highest at `corr=0`. Positive correlation across the trials helps reduce FWER and FDR, without hurting TPR or F1. Negative correlation across the trials decreases TPR and F1 even more so than FWER and FDR.

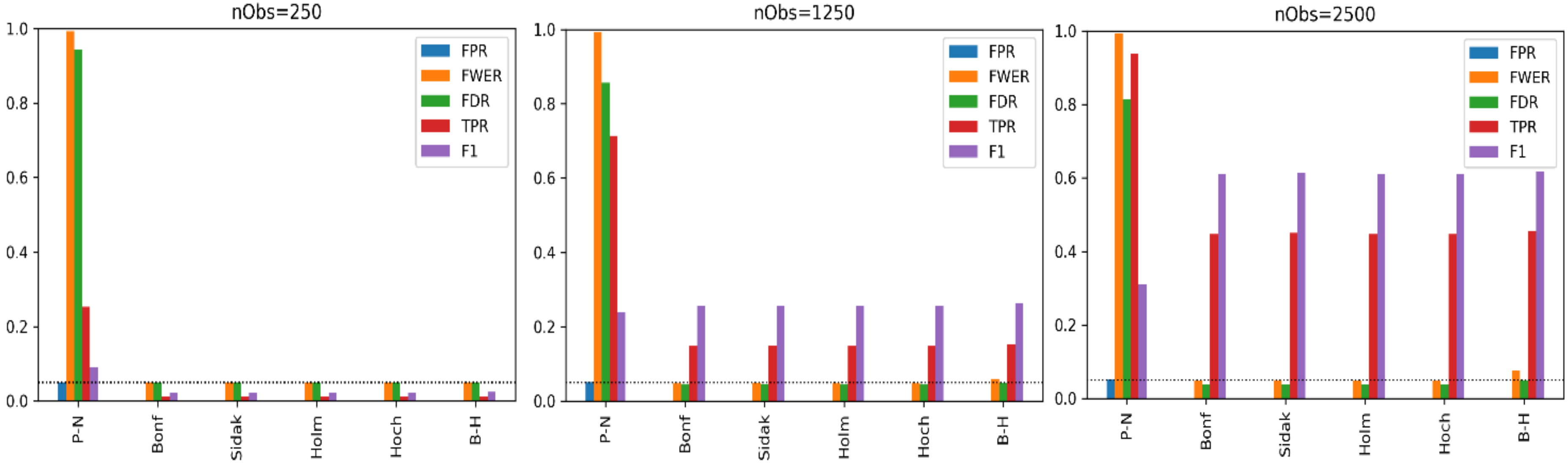
The Pearson-Neyman (P-N) test delivers the targeted FPR of 5%, however its FWER is extremely high, because this test does not control for multiple testing. All other tests perform similarly well in terms of delivering their targeted FWER of 5%. Somewhat disappointingly, Holm, Hochberg (Hoch) and Benjamini-Hochberg (B-H) do not perform materially better than Bonferroni (Bonf) in terms of higher TPR and F1. All methods appear to be acceptable approaches under the correlation conditions of this experiment.

Size Effects



As meanSR increases, the F1 is higher among methods that control for multiple testing (i.e., all but Pearson-Neyman). For meanSR=.03 (equivalent to a Sharpe ratio of 0.5, annualized on a frequency of 250 observations per year), controlling for multiple testing does not improve F1, however achieving Pearson-Neyman's F1 would require implementing every single false positive (which occur in 5% of the negative conditions). When investigating small-size phenomena, it is critical to rely on theoretical explanations, and not only statistical evidence, even after controlling for multiple testing.

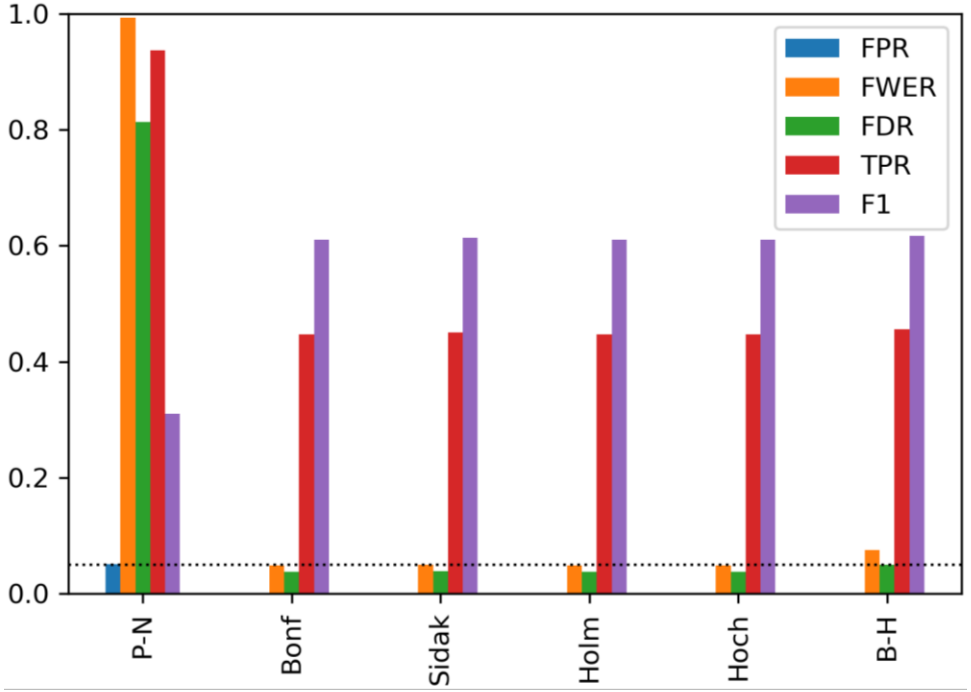
Sample Length Effects



As nObs increases, the F1 is higher among methods that control for multiple testing. For nObs=250 (equivalent to one year of daily observations, excluding weekends), controlling for multiple testing does not improve F1. The 2 or 3 years track record requirement, customary in the financial industry, does not appear to be enough to prevent false discoveries, even after controlling for multiple testing.

Scenario Analysis

- Consider the scenario where
  - `corr=0`
  - `meanSR=1./250**.5`
  - `nObs=2500`
- This scenario is representative of microscopic alpha research projects
- Under those circumstances, methods that control for multiple testing achieve a higher F1 than the standard Pearson-Neyman framework, *without requiring the implementation of a large number of false positives*

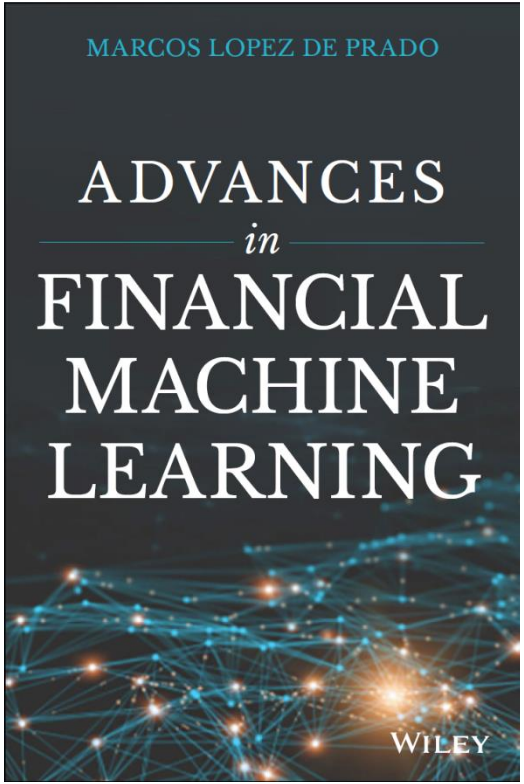


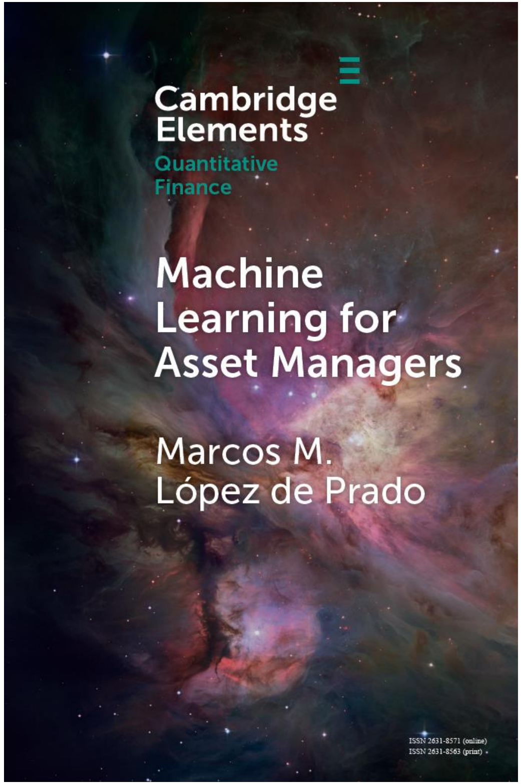
Conclusions

Practical Implications for Strategy Selection

- Most studies conducted in finance (including papers published in the top academic journals)
  - control for FPR (*p*-values) instead of FDR
  - fail to control for multiple testing
- Historical backtests have low power, and can be overfit after a small number of trials, because
  - sample lengths are short
  - variables are highly codependent
  - the unconditional probability of a positive condition (true strategies) is low
- Consequently, most discoveries published in finance are likely false
- **Monte Carlo simulations show that quantitative funds would achieve a higher F1 (hence better performance) if they controlled for FWER and/or FDR**

For Additional Details





The first wave of quantitative innovation in finance was led by Markowitz optimization. Machine Learning is the second wave and it will touch every aspect of finance. López de Prado's Advances in Financial Machine Learning is essential for readers who want to be ahead of the technology rather than being replaced by it.
— Prof. **Campbell Harvey**, Duke University. Former President of the American Finance Association.

Financial problems require very distinct machine learning solutions. Dr. López de Prado's book is the first one to characterize what makes standard machine learning tools fail when applied to the field of finance, and the first one to provide practical solutions to unique challenges faced by asset managers. Everyone who wants to understand the future of finance should read this book.
— Prof. **Frank Fabozzi**, EDHEC Business School. Editor of The Journal of Portfolio Management.

Disclaimer

- The views expressed in this document are the authors' and do not necessarily reflect those of the organizations he is affiliated with.
- No investment decision or particular course of action is recommended by this presentation.
-

www.QuantResearch.org