Forking Paths in Empirical Studies

FDP INSTITUTE® by CAIA Webinar

Welcome

We will begin promptly at 10 AM ET.

If you are unable to hear the speakers, please let us know in the chat box.

You may enter your questions in the Q&A, we will address them at the end of the presentation. You can find a copy of the slide deck and recording of this webinar: www.fdpinstitute.org/webinars

FDP INSTITUTE®

Financial Data Professional Institute

FDP Institute provides world class training and education to financial professionals to meet the accelerating needs of digital transformation in the industry.

Introductions

Hossein Kazemi, PhD, CFA
Senior Advisor,
CAIA Association &
FDP Institute

Guillaume Coqueret
Associate Professor,
Emlyon Business School

Today's Topic:

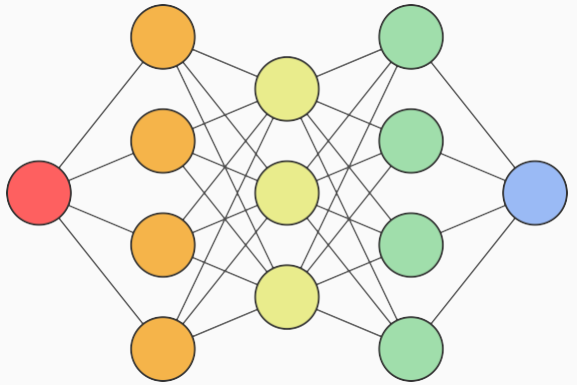
Forking Paths in Empirical Studies

Forking paths in empirical studies

FDP Webinar - 2023-05-23

Guillaume Coqueret (EMLYON Business School)

2022-12-15



Introduction

Starting point: a replication crisis in finance?

THE JOURNAL OF FINANCE • VOL. LXXII, NO. 4 • AUGUST 2017

Presidential Address: The Scientific Outlook in Financial Economics

CAMPBELL R. HARVEY*

Is There a Replication Crisis in Finance?

106 Pages • Posted: 5 Mar 2021 • Last revised: 7 Mar 2022

Theis Ingerslev Jensen
Copenhagen Business School

Bryan T. Kelly
Yale SOM; AQR Capital Management, LLC; National Bureau of Economic Research (NBER)

Lasse Heje Pedersen
AQR Capital Management, LLC; Copenhagen Business School - Department of Finance; New York University (NYU); Centre for Economic Policy Research (CEPR)

THE JOURNAL OF FINANCE • VOL. LXXVI, NO. 5 • OCTOBER 2021

The Limits of *p*-Hacking: Some Thought Experiments

ANDREW Y. CHEN

An ounce of epistemology

the path towards scientific dissemination

it all starts with an **idea**

Empirical phase, with choices (data extraction, missing data, outliers, variables, estimators, etc.)

results are good!

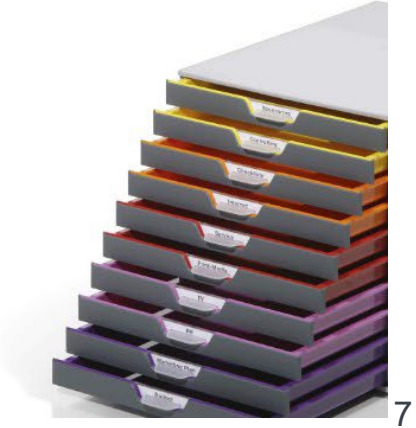
write a paper & communicate results!

results are bad!

same player tries again

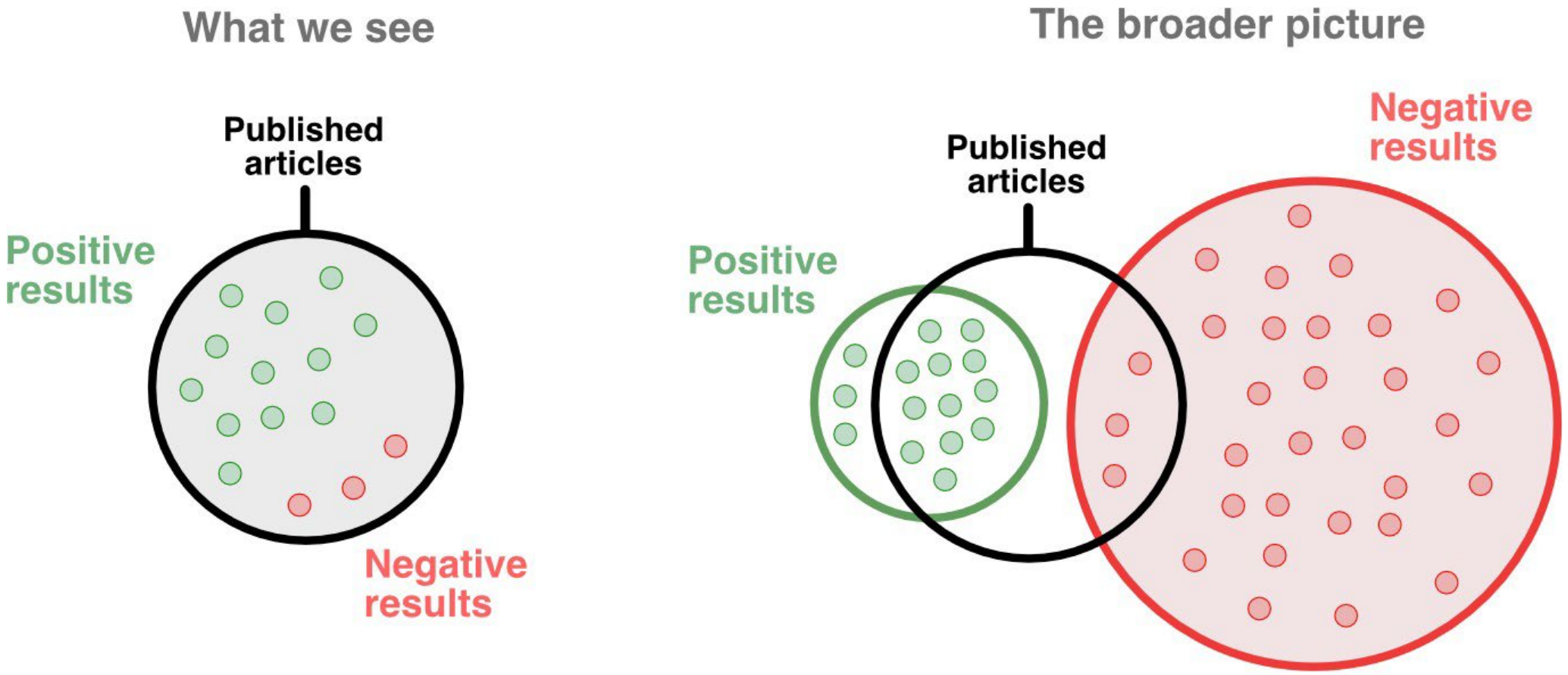
model stage (theory)

results are discarded = file-drawer problem



The outcome

The literature provides a biased picture...



source: Sam Westwood

In short

"*In this garden of* ***forking paths****, whatever route you take seems predetermined, but that's because the choices are done implicitly. The researchers are not trying multiple tests to see which has the best p-value; rather, they are using their scientiic common sense to formulate their hypotheses in reasonable way, given the data they have.* ***The mistake is in thinking that, if the particular path that was chosen yields statistical signiicance, that this is strong evidence in favor of the hypothesis***."

Gelman & **Loken** - *The Statistical Crisis in Science* Am. Sc. (2014)

One solution: exhaustiveness

- Often, papers propose baseline results, followed by (selective?) **robustness checks** (one variable at a time).
- Often, reviewers ask for more checks!
- Sometimes, results cannot be replicated, and they change (qualitatively) when the empirical protocol is slightly altered.
- This leads to *shaky* conclusions and lack of trust.

→ We propose to include **multiple variations** of the initial protocol as the baseline output.

Related literature

Recently, on sensitivity to research design:

- The Influence of Hidden Researcher Decisions in Applied Microeconomics, **Huntington-Klein et al.**, Econ. Inq. 2021
- Methodological variation in empirical corporate finance, **Mitton**, RFS 2022
- Non-standard errors, **Menkveld et al.**, JF 2023
- Computational Reproducibility in Finance: Evidence from 1,000 Tests, **Perignon et al.**, SSRN 2023

On asset pricing factors/anomalies:

- The devil in HML's details, **Asness et al.**, JPM 2013
- Non-Standard Errors in Asset Pricing: Mind Your Sorts, **Soebhag et al.**, SSRN 2022
- Non-Standard Errors in Portfolio Sorts, **Walter et al.**, SSRN 2022

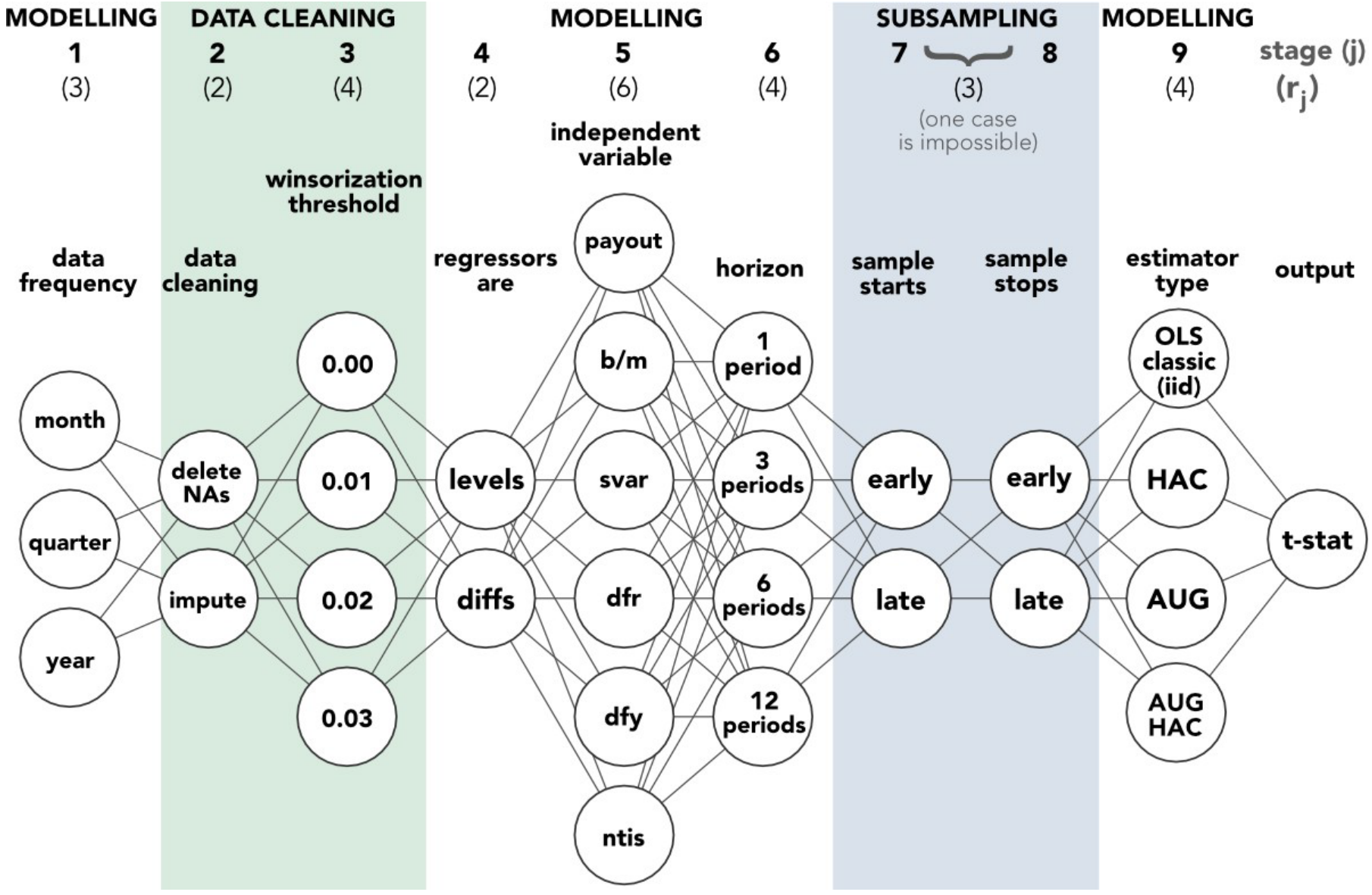
Abstract representation: composition of mappings

- The empirical part of research process starts with some input which we call D (initial version of the data)
- The study is modelled as a sequence of operations f_j so that the reference research output (e.g., one t-statistic) is such that

$$o_J(D) = \left[\bigcirc_{j=J}^{1} f_j\right](D) = f_J \circ f_{J-1} \circ \cdots \circ f_1(D),$$

where $f_j : S_j \mapsto S_{j+1}$, with S_1 and S_{J+1} encompassing the sets of feasible input D and output values, respectively.

Illustration with equity premium prediction



The case of Lipschitz mappings

If we assume that for $D, D \in S_j$, there exists some constant $c_j > 0$ such that

$$\| f_j(D) - f_j(D) \| \leq c_j \| D - D \|,$$

then, for $0 \leq K < J$,

$$\left\| \left[\bigcirc_{j=J}^{K+1} f_j \right](o_K(D)) - \left[\bigcirc_{j=J}^{K+1} f_j \right](o_K(D)) \right\| \leq \| D - D \| \prod_{j=K+1}^{J} c_j.$$

i.e.: there is a **compounding effect** of the number of mappings (more mappings = **larger range of outcomes**, unless they are contracting or non-expanding, which rarely occurs).

So what?

- By generating a large number of alternative outcomes, we get a **more complete picture** of the problem we investigate.
- We can report the **full distribution** of outcomes, and not just a few values that corroborate our priors (and increase odds of publication). → test for p-hacking!
- We can resort to **averaging** and build robust confidence intervals.
- We can determine which particular **design choices** have an impact on the distribution (or mean) of outcomes.
- We can quantify the **widening speed of the range of results** as we add new layers in the protocol. These *hacking intervals* were coined in A theory of statistical inference for ensuring the robustness of scientific results, Coker et al., (2021).

Application n°1: equity premium prediction

The p-curve of predictive regressions (over 13,824 paths)

The data is from Goyal, Welch & Zafirov (2021 follow up from 2008)

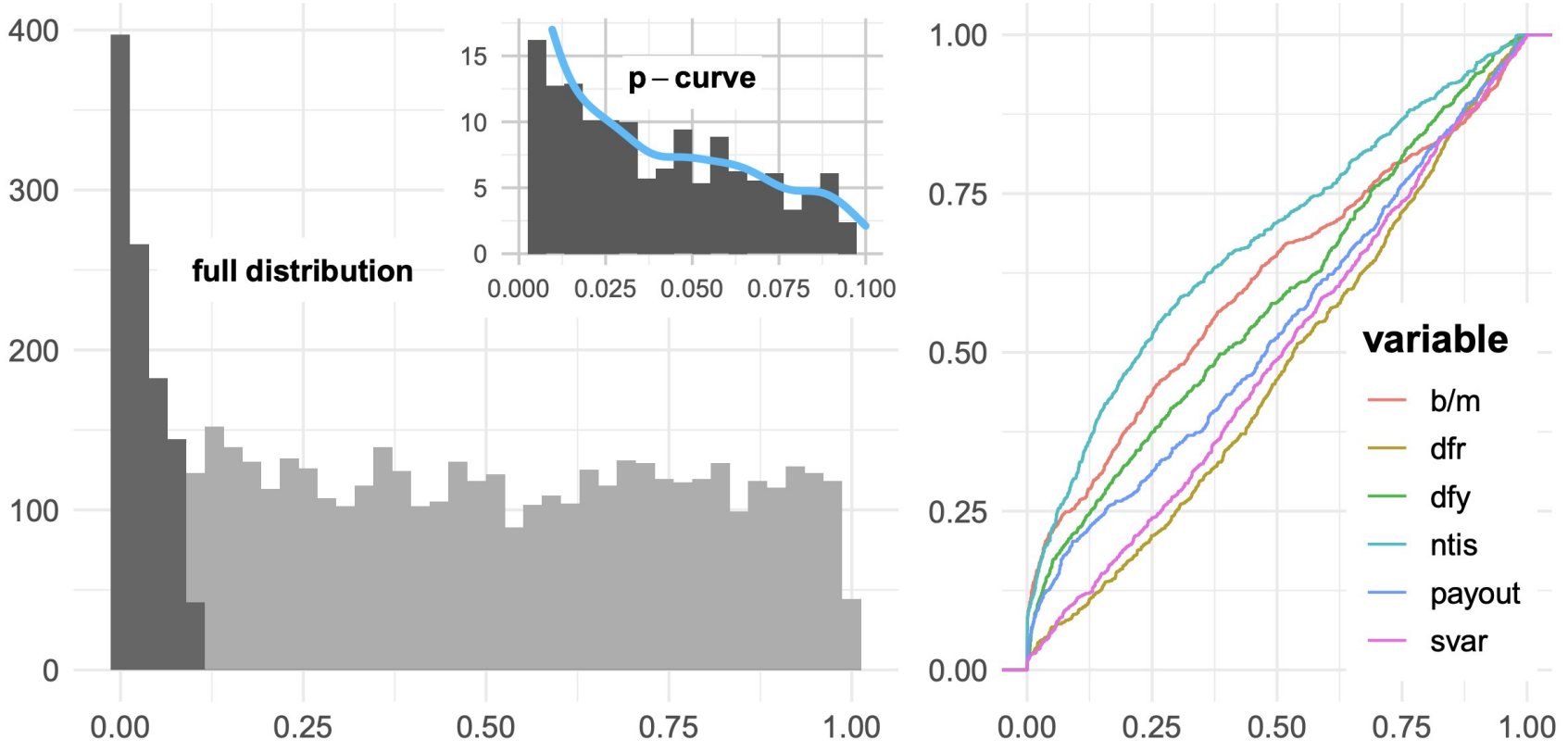


Figure 4: **Distribution of p-values.** In the left panel, we plot the histogram of all p-values, as well as the p-curve (Simonsohn et al. (2014a)), which is the restriction of the distribution to the interval of significant values (which we take to be [0,0.1]). In the right panel, we show the cdf of the p-values, sorted by independent variable. Results for regressions with fewer than 30 observations are discarded.

Impact of mappings (1/2)

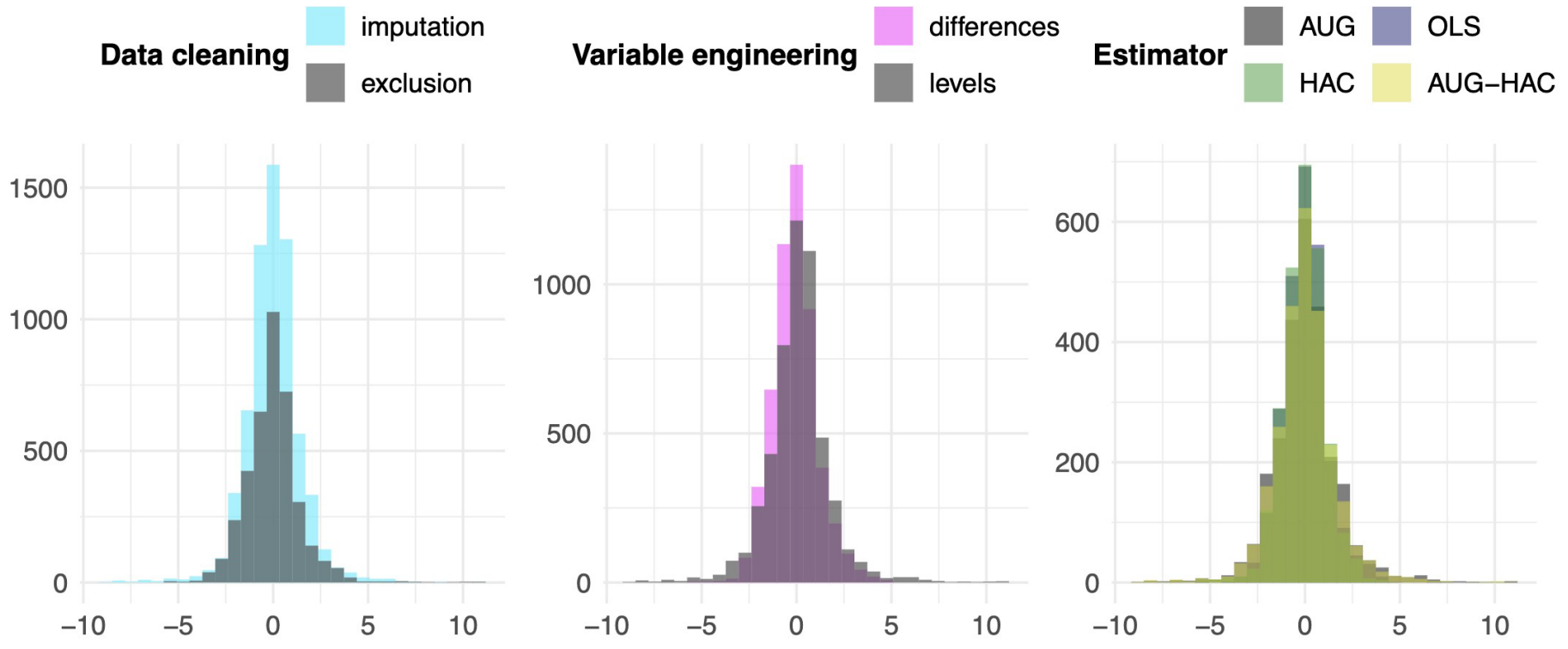
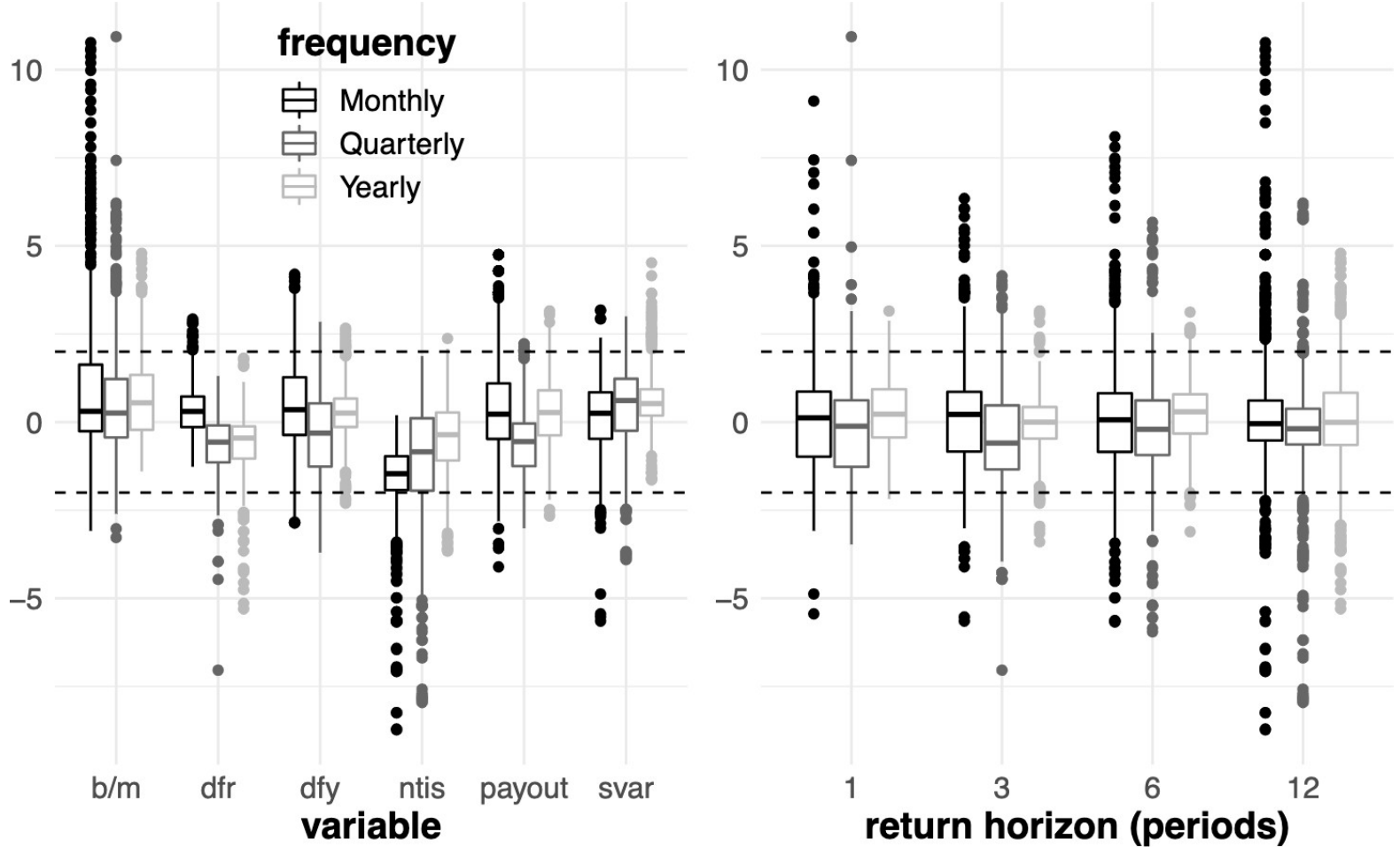


Figure 11: **Impact of mappings: robustness checks.** We report the distribution of t-statistics for two binary choices in mappings, plus the final estimator type. Results for regressions with fewer than 30 observations are discarded.

Impact of mappings (2/2)

The equation is: $r_{t+k} = a + bx_t + e_{t+k}$



Frequentist model averaging

Following Burnham and Anderson (2004)

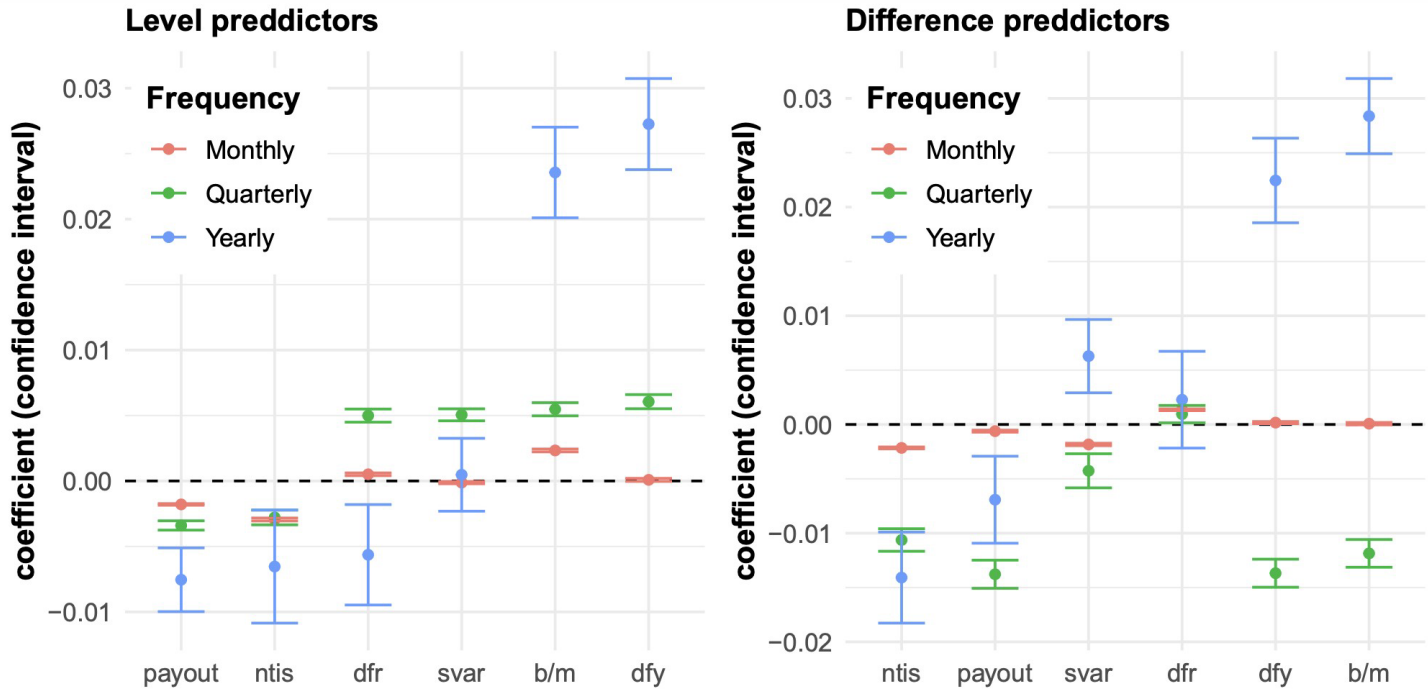


Figure 8: **Frequentist model averaging**. We display average coefficients within their 95% confidence interval. Coefficients stem from Equation (20). Confidence intervals are defined by $[\hat{b}_* - 1.96\sigma_*^2/\sqrt{T_*}, \hat{b}_* + 1.96\sigma_*^2/\sqrt{T_*}]$, where $T_* = \sum_{j=1}^{J} w_j T_j$, with T_j being the sample size of model j. The left panel displays results when predictors are levels, while the right one focuses on differences of variables. To allow comparisons, all predictors are scaled to have unit variance before estimation.

Rate of increase of intervals (1/2)

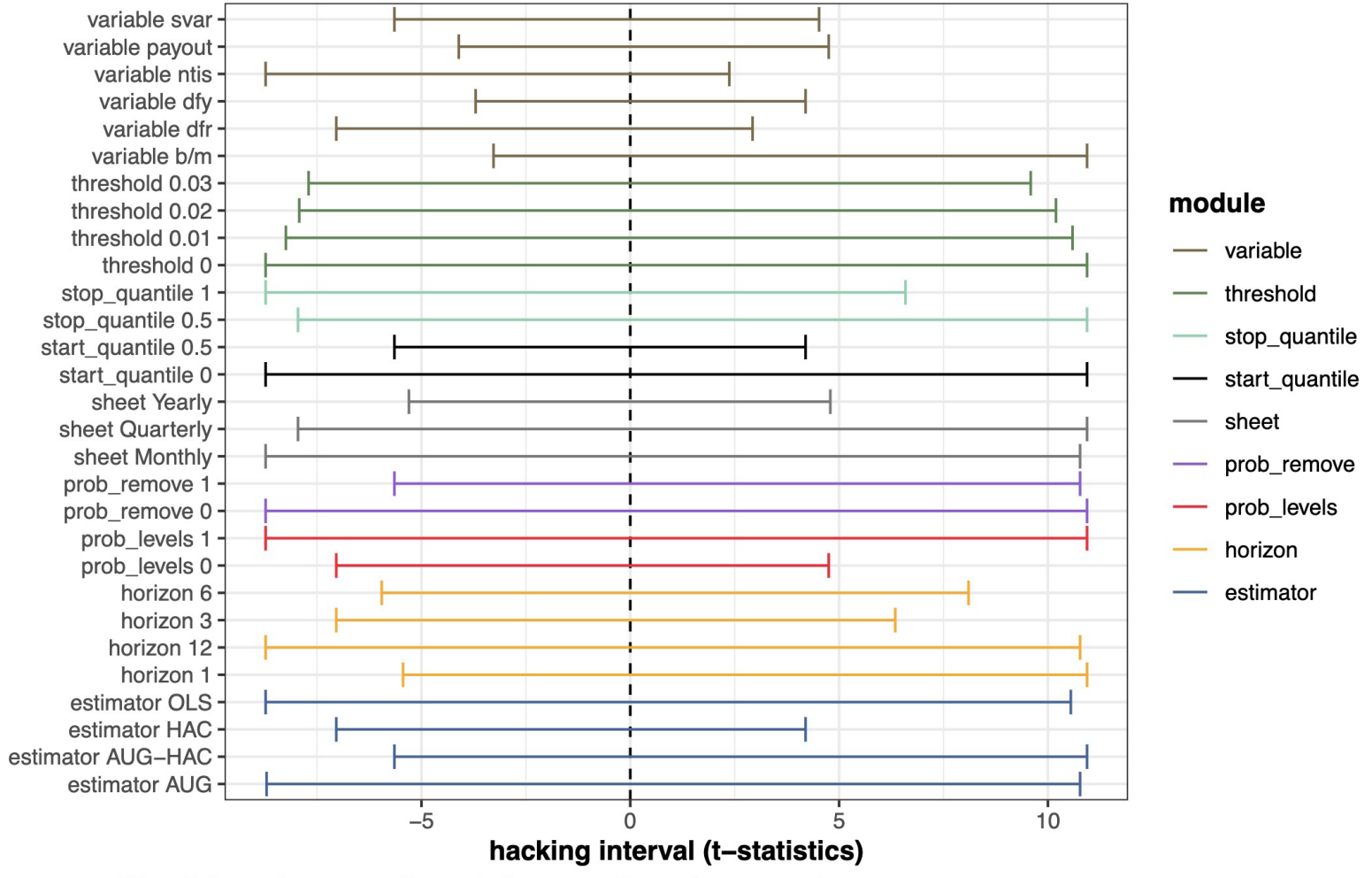
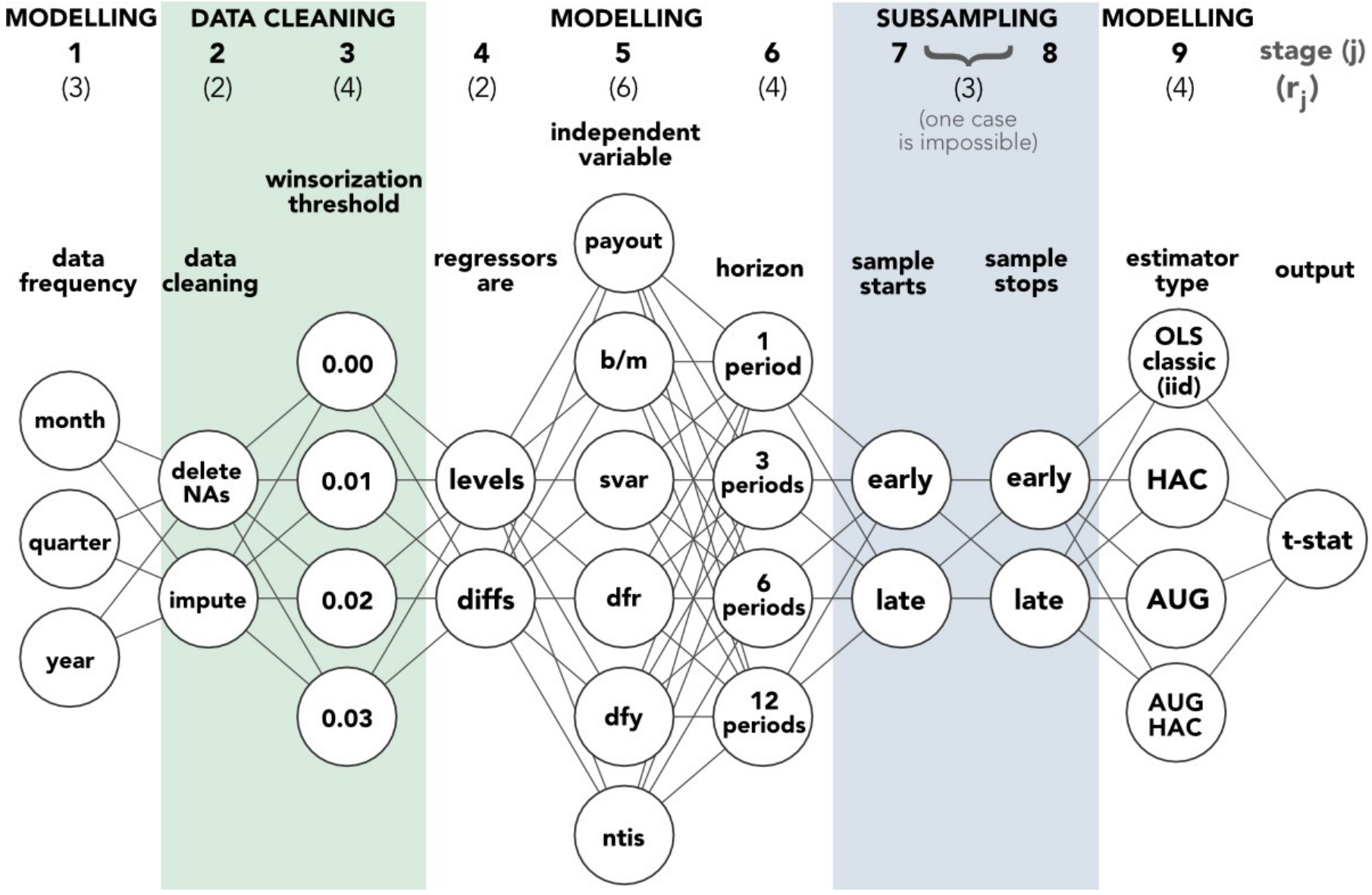
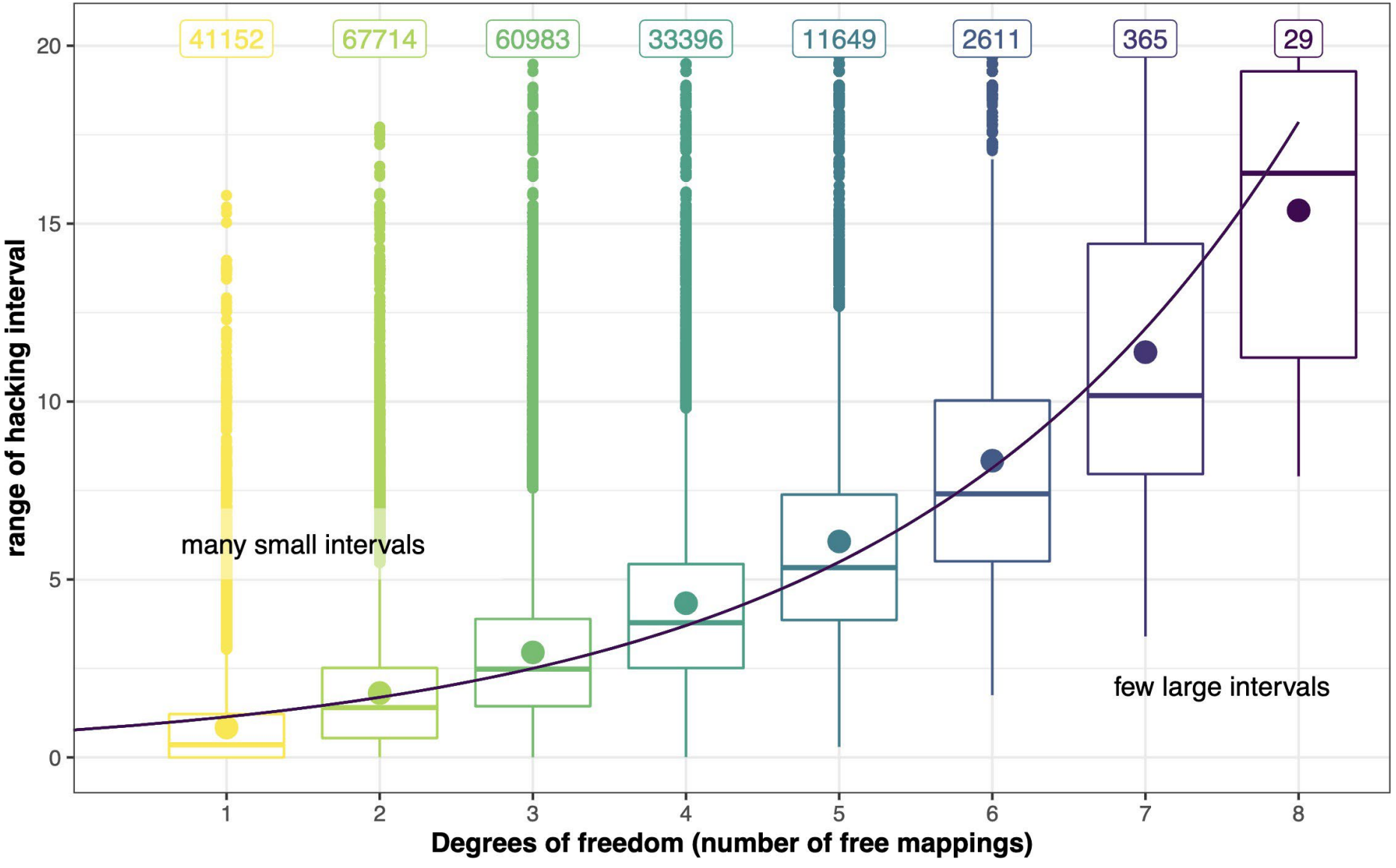


Figure 13: **Hacking intervals with one fixed mapping**. We show the intervals of t-statistics obtained by fixing one mapping. Each option of the mapping is tested and all combinations of all other mappings are spanned to generate the intervals. The nine modules (i.e., mappings) are shown with colors.

Reminder


MODELLING
DATA CLEANING
MODELLING
SUBSAMPLING
MODELLING
1
2
3
4
5
6
7
8
9
stage (j)
(3)
(2)
(4)
(2)
(6)
(4)
(3)
(one case is impossible)
(4)
(r_j)
independent variable
winsorization threshold
data frequency
data cleaning
regressors are
horizon
sample starts
sample stops
estimator type
output
month
quarter
year
delete NAs
impute
0.00
0.01
0.02
0.03
levels
diffs
payout
b/m
svar
dfr
dfy
ntis
1 period
3 periods
6 periods
12 periods
early
late
early
late
OLS classic (iid)
HAC
AUG
AUG HAC
t-stat

Rate of increase of intervals (2/2)



Application n°2: asset pricing anomalies

Context

Since ... and the cross-section of expected returns, it has become customary to **question the validity of factors**.

→ Indeed, why is it so "easy" to find factors, but hard to make money out of them? Some answers in Zeroing in on the Expected Returns of Anomalies (Chen & Velikov JFQA Forthcoming)

As is already shown in The devil in HML's details, small tweaks in the construction of the factors can lead to deterioration in performance. This is risky for investors!

Intuitively, they seek a construction of factors that will exhibit a performance that is not sensitive to implementation details.

Introducing exhaustive multiple testing (EMT)

Combining two types of approaches.

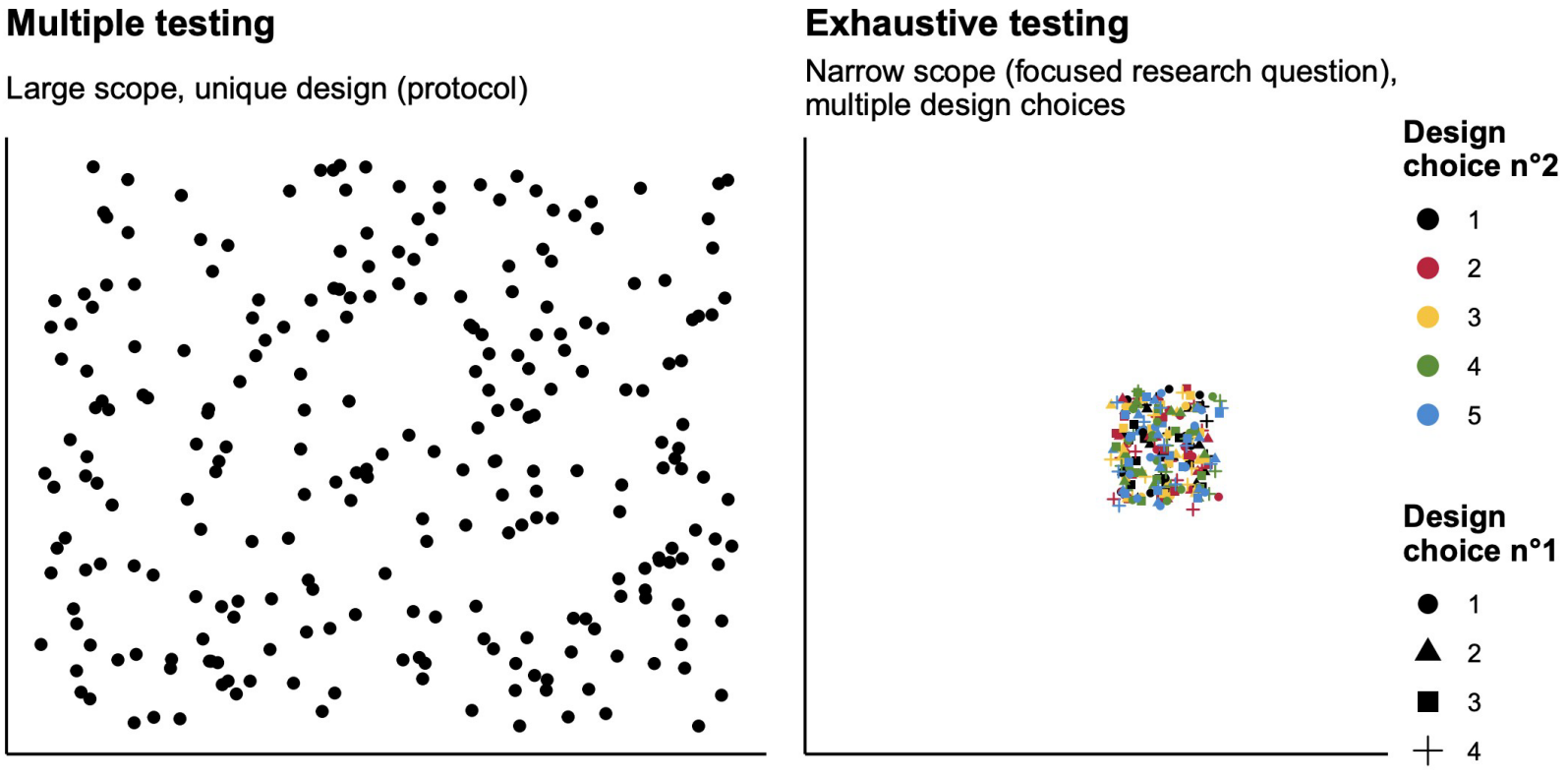


Figure 1: Illustration of **model rich and protocol poor** versus **model poor and protocol rich** studies.

Formally (1/2)

We follow the **BRC** version of An Evaluation of Alternative Multiple Testing Methods for Finance Applications (Harvey & Liu RF 2020)

We are given $T \times N$ observations $x_{t,n}$, where T is the sample size and N the number of tests. These observations are bootstrapped B times to yield a $B \times T \times N$ tensor $x^{(b)}_{t,n}$. Here, b is the index of the bootstrapped sample.

Bootstrapped statistics are

$$t_n^{(b)} = \sqrt{T}\,\frac{\mu_n^{(b)} - \mu_n}{\sigma_n^{(b)}},$$

where $\mu_n^{(b)}$, $\sigma_n^{(b)}$ are the sample mean and standard deviation of each bootstrap series. μ_n is the sample mean of the original (non bootstrapped) data.

Formally (2/2)

We write $\tilde{t}^{b}_{n}$ for the statistics ordered such that $\tilde{t}^{(b)}_{n} \geq \tilde{t}^{(b)}_{n+1}$, so that, for each bootstrap sample b, $\tilde{t}^{(b)}_{1}$ is the largest statistic. We are then given a confidence level l, say $l = 95\%$. The target threshold for the test is then the l quantile of the vector $\tilde{t}^{(b)}_{1}$.

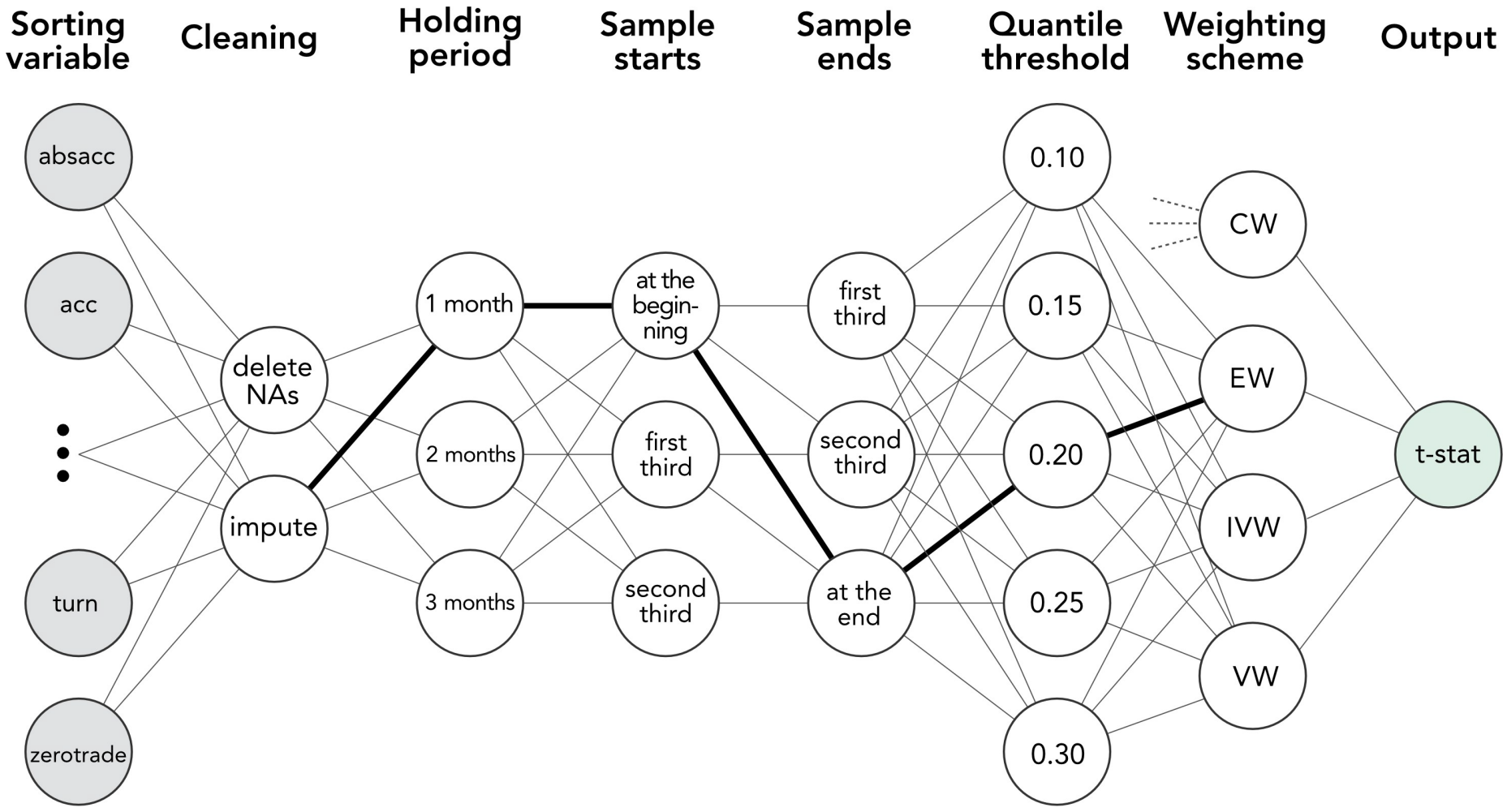
What we refer to **exhaustive multiple testing** is replacing bootstrapping by forking paths:

$$t^{(p)}_{n} = \sqrt{T}\,\frac{\mu^{(p)}_{n} - \mu_{n}}{\sigma^{(p)}_{n}},$$

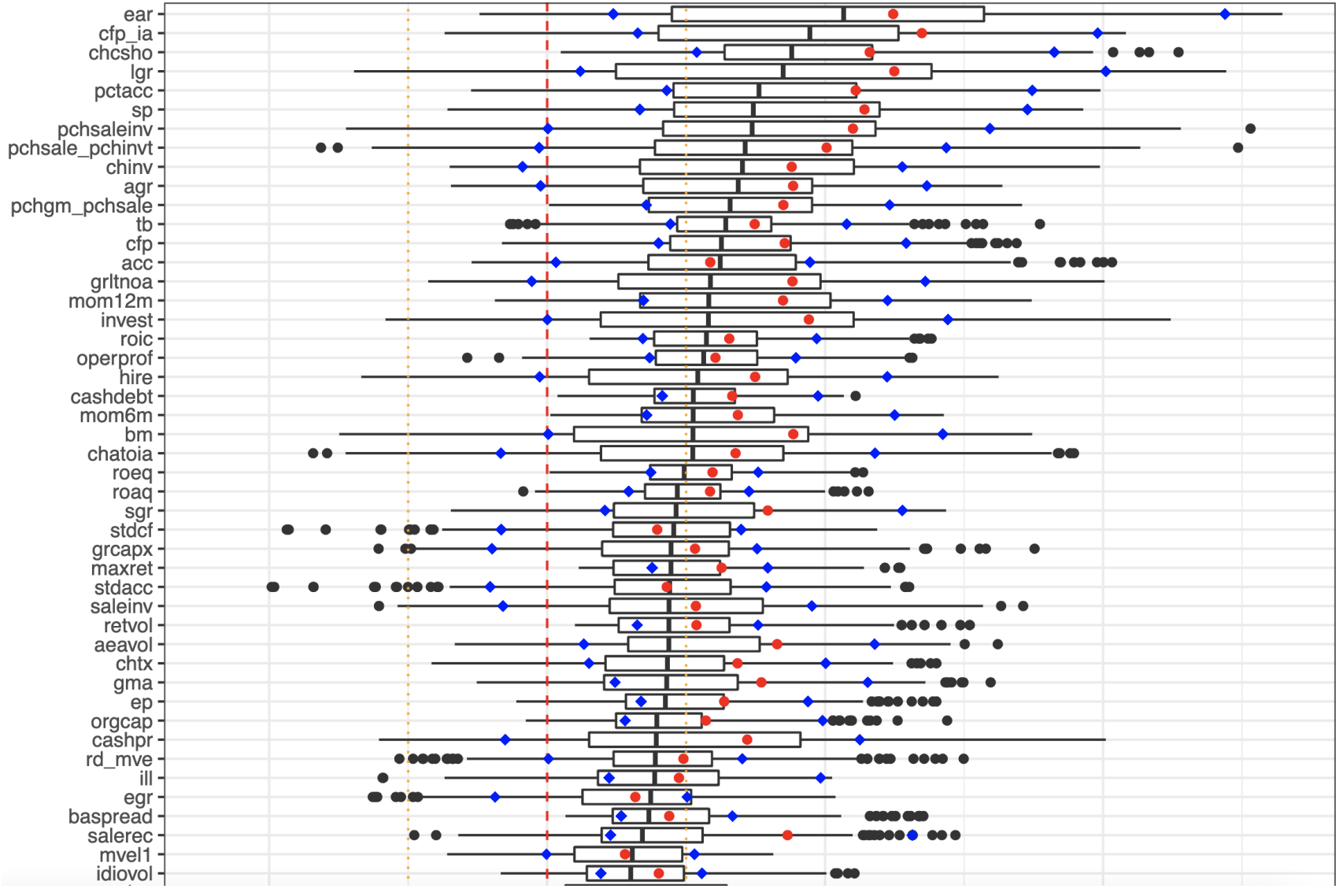
and **processing the corresponding thresholds**.

EMT for asset pricing anomalies: paths

The data is obtained from Dacheng Xiu's website, from the Empirical Asset Pricing paper.



EMT for asset pricing anomalies: results



EMT for asset pricing anomalies: statistics

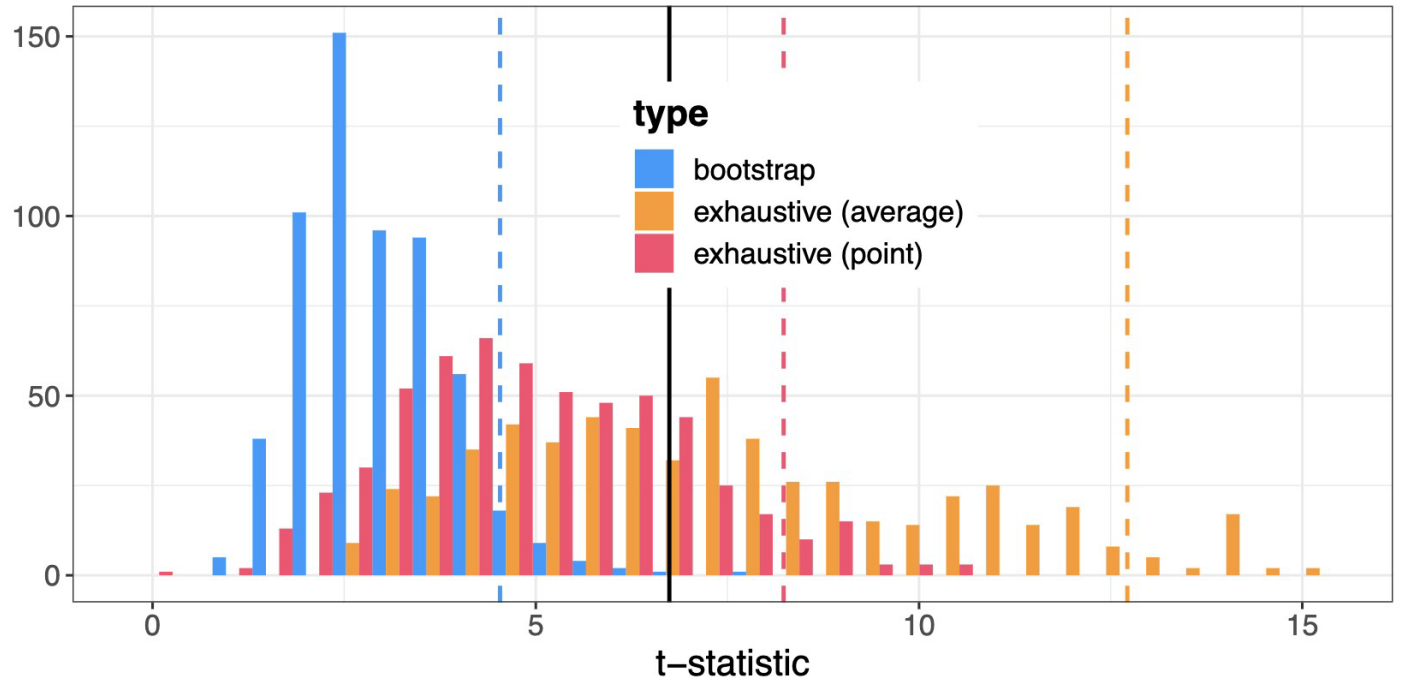
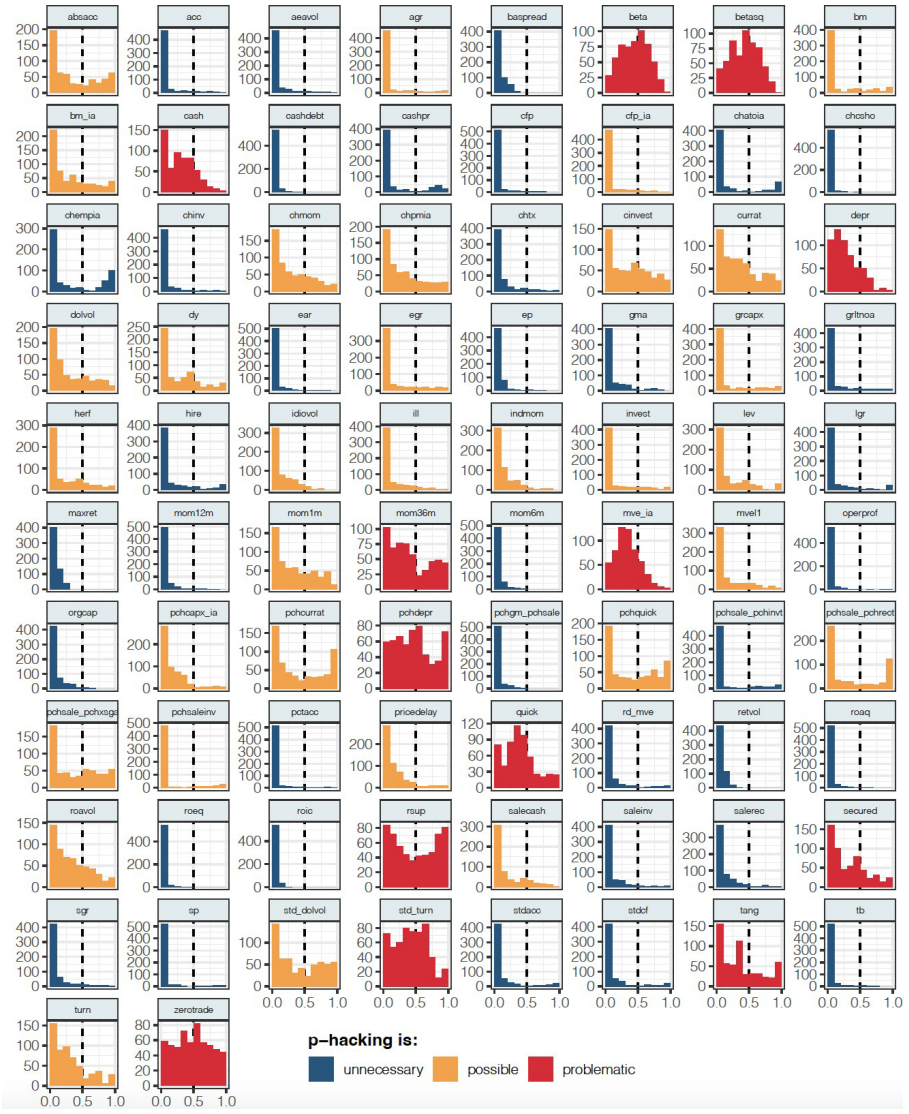


Figure 11: **Distribution of bootstrapped and paths-related maximum statistics.** We produce the histogram of the maximum statistics stemming from bootstrapping ($\tilde{t}_1^{(b)}$, in blue) and forking paths ($\tilde{t}_1^{(p)}$, in orange and red). The sequences are derived from equations (14) and (15), respectively. The vertical black line is the benchmark t-statistic of the *best* anomaly (*cfp_ia* in this case) for the default path. The vertical dotted lines correspond to the 95% quantile of the maximum statistics of each type. The difference from the two exhaustive distributions comes from the benchmark μ_n used to compute the statistics. The point-wise values are obtained when μ_n is the average anomaly return of the default path described above. The average values correspond to the case when μ_n is the average of factors' returns over all paths.

A word of caution

- Modern tools for **multiple testing** are aimed at controlling the odds of **false negatives** (when effects are unduly discarded as insignificant).
- Basically, they want to be good at detecting both **error types** (better overall accuracy)!
- Here, we do not do that, at all. We posit that asset managers have asymmetric preferences and put much more emphasis on false positives (investments that disappoint) than on false negatives (missed opportunities).
- In short, our method aims to single out the strategies (factors) that perform well across **almost all specifications**.

Final round: p-hacking detection



[Detecting p-hacking]

(Elliott et al., ECTA 2022):

the p-curve should be completely monotone on (0,1/2).

Ongoing work

Environments

Inspired from Causal inference by using invariant prediction: Identification and confidence intervals.

Let us consider the case of a simple model $y = Xb + e$ in which b is random. We are given the opportunity to estimate this model from many *pseudo* environments, which are couples (y^p, X^p), so that estimates depend on these environments $\hat{b}^p$.

Intuitively, if we span a large number of environments, we should hope that the empirical cdf of the $\hat{b}^p$ converges to that of b. The challenge is to devise a theoretical framework in which this can occur $\rightarrow$ not so simple!

A try

On way to do so is to assume that $\hat{b}^p$ are random variables which have the same distribution as b (strong assumption). **Randomness comes from samples**, not errors (at least not directly).

The major issue is then: how can we characterize and handle the correlation between the outcomes from the paths. Indeed, a large majority of paths are relatively close, so that their outcomes should be **non-negligibly correlated**.

Interesting directions

The Empirical Distribution of a Large Number of Correlated Normal Variables,

Azriel & Schwartzman, JASA 2015

Z_p are $N(0, 1)$ gaussian variables, Σ_P the covariance matrix of $(Z_1, \ldots, Z_P)$ and $\hat{F}_P(z)$ the empirical cdf. If

1. $||\Sigma_P||_{1,2} \overset{P\to\infty}{\to} 0$, then $\sup_z E[(\hat{F}_P(z) - \Phi(z))^2] \to 0$.
2. Otherwise, $E[(\hat{F}_P(z) - \Phi(z))^2]$ does not converge to zero for any z.

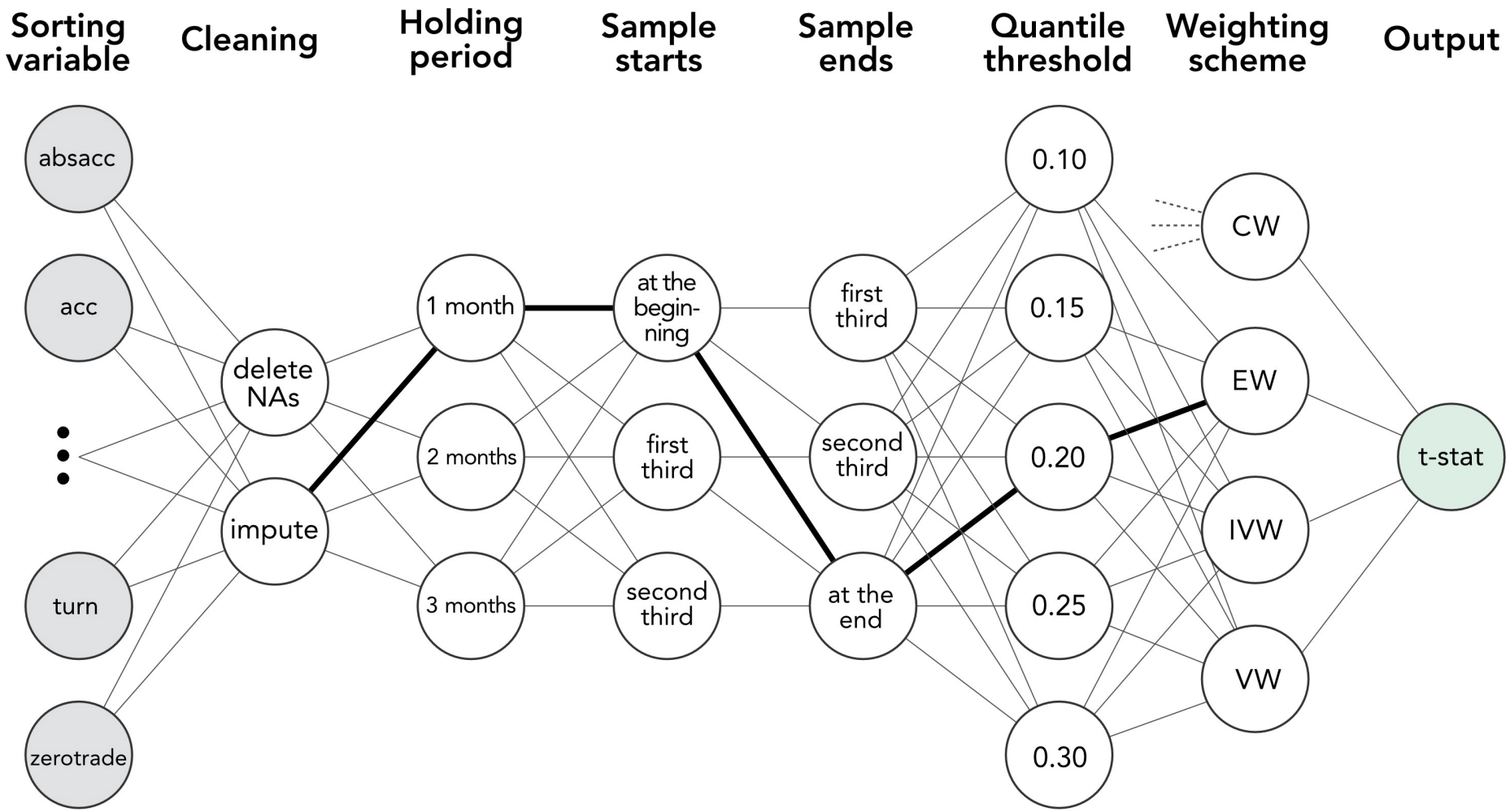
Also: Concentration inequalities for empirical processes of linear time series,

Chen & Wu, JMLR 2018 → importance of memory & tails in MA processes.

Correlation via proximity

- The critical object is the **correlation matrix** of the outputs from the paths. What can assume about it?
- Are its elements all positive...?
- Heuristically, it would make sense that the correlation between two paths increases with the **proximity** between these paths. How can we define proximity?

Illustration: back to paths



Some notations

- Mappings (layers of decision) f_j have r_j deterministic options which the researcher must choose from, and which we write $f_{j,r}$, for $r = 1, \dots r_j$, where $r_j \geq 2$. Recall: $P = \prod_{j=1}^{J} r_j$. Hence a path is just the collection of choices $p := (f_{j,r_{p,j}})_{1 \leq j \leq J}$. Simpler notation: $f_{j,r(p)}$.
- For each layer, there is a distance function that measures the proximity between 2 options: $D_j(f_{j,r(p)}, f_{j,r(q)})$.
- We aggregate them across layers to obtain the distance between 2 paths: $D(p, q) = \sum_{j=1}^{J} \omega_j D_j(f_{j,r(p)}, f_{j,r(q)})$.

A simplification

Henceforth, we set $D(p,q) = \#\{l,\, f_{l,r(p)} \neq f_{l,r(q)}\}$, i.e., the number of choices that differ from *p* to *q*. Then for any path *p*, the number of other paths which have an arbitrary distance of *d* (with *p*) is

$$\sum_{n=1}^{\binom{J}{d}} \prod_{s=1}^{d} (r_{j_{s,n}} - 1) = e_d(r_1 - 1, \ldots, r_J - 1), \qquad d \geq 1,$$

where

$$e_k(x_1, \ldots, x_J) = \sum_{1 \leq j_1 < \cdots < j_k \leq J} x_{j_1} \ldots x_{j_k}$$

are elementary symmetric polynomials.

Two choices

In order to generate an increasing number of paths, we can:

- augment the number of **layers** J, with finite number of options r_j; or
- let some of the **sets of options** increase indefinitely (e.g.: continuous thresholds, subsamples), but for a finite J.

Or let both J and some $r_j \to \infty$.

This can matter a lot!

A toy result and extensions

If $\rho(p,q) = \rho^{D(p,q)}$ for some $\rho \in (0,1)$, then

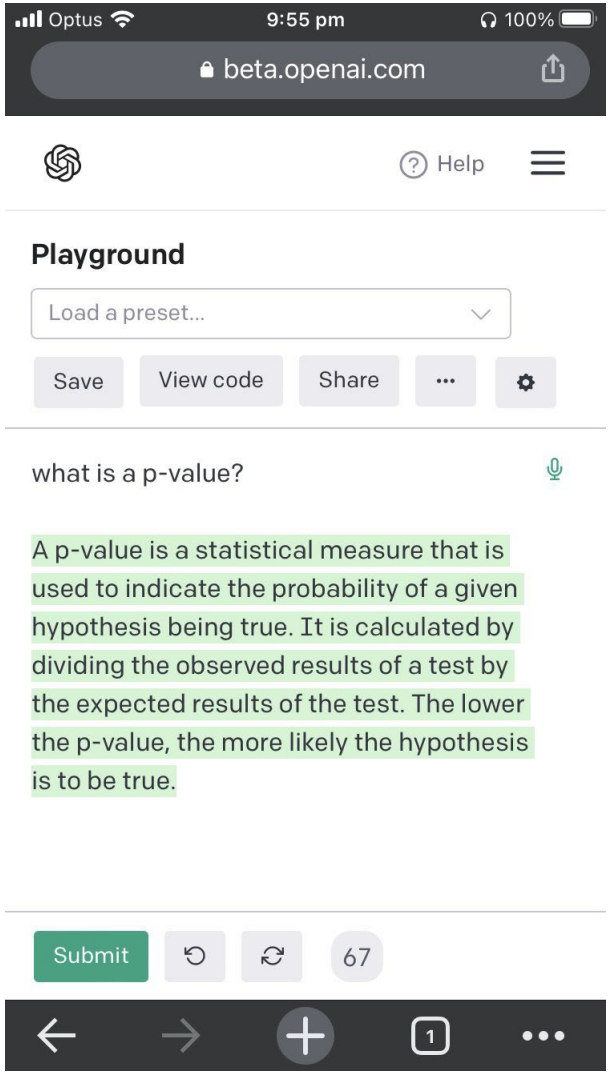
$$P^{-2} \sum_{1 \le p,q \le P} \rho(p,q) \overset{J \to \infty}{\longrightarrow} 0.$$

But! The term does not converge to zero when J is fixed and some $r_j \to \infty$.

Extensions include:

- Other distances: **M-dependent** paths
- More general definition of **distance** between paths
- Procedure to estimate/quantify correlations

The holy grail...?



1. Redefining p-values: $P[b > 0|Data]$... Out of reach for now...
2. Link with the **empirical null distribution**:
   - False discovery rate control with unknown null distribution Roquain & Verzelen (Ann. Stat. 2022)
   - Semi-supervised multiple testing Mary & Roquain (EJS 2022)

Conclusion

Takeaways (& limitations)

- **Main message**: exhaustive protocols and reporting of outcomes should become the baseline results!
- But of course, there is one important limitation: this is only possible when 1 baseline result takes a reasonable amount of time & computing power.
- And this is a tough sell because it requires more efforts from researchers...

Q&A

Please join us for our upcoming webinar:

Register Here: https://bit.ly/3mIKE7x

Thank You

FDP INSTITUTE®
by CAIA

Contact Us:

fdpinstitute.org

info@fdpinstitute.org

@FDPInstitute

linkedin.com/company/FDP Institute