Forking paths in financial economics

Guillaume Coqueret

Forking paths in financial economics

Guillaume Coqueret*

December 6, 2023

Abstract

We argue that spanning large numbers of degrees of freedom in empirical analysis allows better characterizations of effects and thus improves the trustworthiness of conclusions. Our ideas are illustrated in three studies: equity premium prediction, asset pricing anomalies and risk premia estimation. In the first, we find that each additional degree of freedom in the protocol expands the average range of t-statistics by *at least* 30%. In the second, we show that resorting to forking paths instead of bootstrapping in multiple testing raises the bar of significance for anomalies: at the 5% confidence level, the threshold for bootstrapped statistics is 4.5, whereas with paths, it is at least 8.2, a bar much higher than those currently used in the literature. In our third application, we reveal the importance of particular steps in the estimation of premia. In addition, we use paths to confirm prior findings in the three topics. We document heterogeneity in our ability to corroborate prior studies: some conclusions seem robust, others do not align with the paths we were able to generate.

1 Introduction

"*Because the empirical economist must deal with nature in all her complexity, it is optimistic in the extreme to hope or believe that standard parametric economic models or probability models are sufficiently adequate to capture this complexity.*" - White (1996)

1.1 Replication: a necessity and a challenge

Empirical studies are built on many choices. Recently, five separate studies[1] have revealed that, given the same replication task, independent researchers can reach vastly different conclusions, including on trivial items such as sample sizes. Such divergences are the consequence of the large

*EMLYON Business School, 23 avenue Guy de Collongue, 69130 Ecully, FRANCE. coqueret@em-lyon.com. I thank Jonathan Lewellen, David Sraer, Olivier Scaillet, Fabio Trojani, Stéphane Guerrier, Aurélien Baillon, Hossein Kazemi, Tony Guida, Gaetan Bakalli and Jean-Yves Gnabo for their constructive feedback. I am also grateful for comments from particpants from the 2022 AFFI conference, as well as the Financial Data Professional Institute, the Geneva Finance Research Institute and the Research Center for Statistics (UNIGE) seminars. Lastly, I am indebted to Andrew Y. Chen for providing data on published asset pricing anomalies.

[1]See Huntington-Klein et al. (2021), Breznau et al. (2022), Gould et al. (2023), Huber et al. (2023) and the fincap study (exploited in Menkveld et al. (2023) and Pérignon et al. (2023)). Another enlightening exercise is provided by FivethirtyEight, in its *Science is not broken* article, in the political science discipline.

number of design options that are left to the appreciation of the empiricist. Examples of these choices are for instance listed in Mitton (2022) for the field of corporate finance.

In the best case scenario, minor shifts will lead to small adjustments, and solid conclusions will not be altered. Nevertheless, sometimes, articles may be retracted because the results are contradicted or simply not robust enough (Rampini et al. (2021) and Boissel and Matray (2022) are two recent occurrences).[2] The replication of results has become a strong imperative in modern research,[3] so that data and code sharing policies have for instance been enforced in most major journals in finance and economics.[4] When results can only be partly replicated (e.g., as in Hou et al. (2020) in asset pricing or Kapoor and Narayanan (2022) in applied machine learning), debates can be fostered to determine if initial conclusions remain valid.

The sensitivity of results to design choices is well-known in the research community, which is why many papers incorporate robustness checks in which analyses are replicated with small alterations in the original empirical protocol. For instance, researchers can reproduce their results on sub-samples, or with alternative estimators, or when testing different values in parametric methods and techniques. They can also evaluate sensitivity measures of estimated parameters (Leamer (1985), Jørgensen (2023)).

Given the publication bias towards positive results,[5] authors can be, knowingly or not, incentivized to produce low p-values in order to demonstrate that their results are *significant*, and hence, worthy of publication.[6] Equipped with a large palette of adjustments (i.e., leeway) in the research design, researchers may be tempted to mostly report those that confirm their priors or theoretical predictions. Indeed, positive results are easier to push forward and increase the odds of acceptance by reviewers and journal editors. From a scientific standpoint, this is suboptimal, because

[2]Retractions are closely followed by a handful of researchers (https://retractionwatch.com) who have compiled and continuously maintain a database of scientific articles that have been retracted from their journals: http://retractiondatabase.org. See also Shepperd and Yousefi (2023) for an analysis of retractions in computer science.

[3]We point to Boylan (2016), Duvendack et al. (2017), Christensen and Miguel (2018), Mueller-Langer et al. (2019), Vilhuber (2020), Colliard et al. (2022), Hofman et al. (2021), Cortina et al. (2022), Peters et al. (2023) and Vu (2022) for discussions on replication and reasons that explain why it may be hard. Recent initiatives are the Institute for Replication in economics and the Open Marketing platform. One of the first attempts of the former pointed towards inflated conclusions (Kjelsrud et al. (2023)).

[4]In **finance**, we can list: Journal of Finance, Review of Financial Studies, Journal of Financial Economics, and Review of Finance. In **economics**, there are at least the following: Econometrica, journals of the American Economic Association (including the American Economic Review), Quarterly Journal of Economics, Review of Economic Studies. In parallel, news services have emerged that propose framework to ease reproductibility, e.g., nuvolos or cascad.

[5]This phenomenon, also known as the *file drawer problem*, has been widely documented in many fields, especially **psychology** (Rosenthal (1979), Collaboration (2015), Stroebe (2019)), **medicine** (Dickersin et al. (1987), Begg and Berlin (1988), Olson et al. (2002), Barnett and Wren (2019) and Van Aert et al. (2019) to cite a few references), **economics** (Leamer and Leonard (1983), De Long and Lang (1992), Stanley (2005), Doucouliagos and Stanley (2009), Doucouliagos and Stanley (2013), Brodeur et al. (2016), Camerer et al. (2016), Ioannidis et al. (2017), Brodeur et al. (2020) and Kasy (2021)), **finance** (Lo and MacKinlay (1990), Harvey (2017), Morey and Yadav (2018) and Harvey and Liu (2021b)), and **accounting** (Chang et al. (2023)). In a recent experimental study encompassing 500 researchers, Chopra et al. (2022) conclude: "*we document that studies with a null result are perceived to be less likely to be published, of lower quality, and of lower importance than studies with statistically significant results even when holding constant all other study features*". Moreover, Serra-Garcia and Gneezy (2021) find that papers that are hard or impossible to replicate are more cited than more transparent studies. Other related issues are bias in research (Fanelli et al. (2017)) and conventionality, whereby mainstream studies with expected outcomes have a higher probability of being published in top journals (Dai et al. (2023)).

[6]The topic of p-hacking is now widely documented in many fields since the seminal work of Sterling (1959), and we point to a few articles on the matter, among many others: Head et al. (2015), Christensen and Miguel (2018) and Brodeur et al. (2022).

failure to reject the null is often informative (Abadie (2020)), as it signals to other researchers where *not* to look. That being said, Blanco-Perez and Brodeur (2020) have shown that editors have the power to change this habit in the refereeing community.

The troubles of false discoveries are also potent and problematic outside academia. In the money management industry, quantitative researchers compete to discover profitable trading strategies. Unfortunately, many of the latter which perform well in-sample end up disappointing out-of-sample in practice (see Bailey et al. (2014), De Prado (2018), Chen and Velikov (2023)). This has spurred a debate on whether there is a crisis of reproducibility in the field.[7] As a solution, many scientists argue in favor of redefining the notion of *statistical significance* (Harvey (2017), Benjamin et al. (2018)), or even propose abandoning it purely and simply,[8] mainly for two reasons. First, because people sometimes confuse p-values with the probability of the null being true, given the data. And second, because if p-values are no longer the ultimate yardstick of scientific discovery, researchers will be less inclined to tilt their empirical protocols so as to produce the sought output (i.e., p-values below some chosen threshold, usually 1% or 5%). Without p-values, no more p-hacking.

The research community has also proposed other avenues to tackle these issues, e.g., by identification and detection methods for false discoveries and p-hacking (Simonsohn et al. (2014a,b), Elliott et al. (2022)), and even solutions via model averaging (Moral-Benito (2015), Steel (2020)), extreme bound analysis (Leamer and Leonard (1983), Granger and Uhlig (1990)), correction measures (Andrews and Kasy (2019)), shrinkage (van Zwet and Cator (2021)), robust aggregation (Rytchkov and Zhong (2020)), noise dissemination (Echenique and He (2023)), new critical values (McCloskey and Michaillat (2023)), specification curve analysis (Simonsohn et al. (2020)), or Bayesian publication decisions (Frankel and Kasy (2022)).

In the present paper, similarly to Fabozzi and de Prado (2018), we argue that one solution, albeit a costly one, is to report the outcomes of a large number of forking paths in empirical studies. Large scale experiments are for example also advised in Milkman et al. (2021) in the field of behavioral science. This is likely to dramatically increase the transparency of the research process and to strengthen the robustness of findings. In a similar spirit, Coker et al. (2021) propose to report "*p-hacking intervals*", which correspond to the total range of possible outcomes (e.g., coefficients, t-statistics, R^2, etc.) that are obtained when spanning sets of hyper-parameters in supervised learning models.

1.2 Beyond replication: confirmation

Being able to replicate findings is an important and necessary step, but it does not ensure the generalization of a result to different empirical settings. When a result is published, its long-term validity will depend on whether it is *confirmed* subsequently by studies which will seek to check if it holds for other data sources, in other geographical zones, or over alternative periods. If it does, then the result can be *generalized* and sees its practical reach extended. Pérignon et al. (2023) put forward two types of reanalyses: reproductions and replications. Reproduction occurs when the dataset is exactly the same as the original study, but the code can be different. Replication signals more leeway: the code and method can be different, or the data and code can also change. When the data, code and method change, the authors still talk of replication, but in the present paper, we will put forward the notion of *confirmation*.

[7] See Bailey and Lopez de Prado (2021), Chen (2021), Harvey and Liu (2021b), Harvey (2021), Chen and Zimmermann (2022b) and Chen et al. (2023). This debate is also ongoing in the medical sciences (see Ioannidis (2005) and Leek and Jager (2017), as well as the discussion in Fanelli (2018))

[8] Among many others, we can point to: Carver (1978), Ziliak and McCloskey (2008), Woolston (2015), Amrhein et al. (2019), McShane et al. (2019), Wasserstein et al. (2019). A more nuanced take is given in Imbens (2021).

Let us exemplify our argument with a well-documented phenomenon, the (cross-sectional) momentum in international markets. The idea is attributed to Jegadeesh and Titman (1993) but has since then been corroborated by many studies. The dimensions along which momentum has been confirmed are numerous, including geographical (Rouwenhorst (1998), Chan et al. (2000), Griffin et al. (2003), Bhojraj and Swaminathan (2006)), chronological (Smith and Timmermann (2022)), industrial (Moskowitz and Grinblatt (1999)), and across asset classes (Asness et al. (2013)) - to cite but a few. This accumulation of evidence, in spite of periodic crashes (Barroso and Santa-Clara (2015), Daniel and Moskowitz (2016)), establishes momentum as a multiply verified pattern in financial economics. This leads us to propose in Table 1 a classification of empirical conclusions, from the least reliable to the most robust evidence.[9] The present paper focuses on the third type for which one contribution is able to span a large number of cases.

Evidence type	Protocols	Publications	Dimensions	Evidence strength
Anecdote	one	one	one	none
Robustness checks	few	one	few	weak
Internal paths	many	one	many	sustained
External confirmation	many	many	one	sustained
Exhaustive documentation	many	many	many	strong

Table 1: **Classification of empirical conclusions**. The number of protocols (i.e., paths) can come either from several publications that treat the same topic, or, as in the present paper, from one publication which generates and reports many results internally. By "*few*", we mean a dozen at most, whereas by "*many*", we mean a hundred at least.

The generation of multiple outputs, based on small variations of similar datasets, shares some similarities with re-sampling techniques, as well as data augmentation and bagging, all of which are sometimes used in machine learning. The premise is that a model which relies on a diversified set of sub-models will benefit from a *wisdom of the crowds* effect, as long as each individual model is relevant (loosely speaking) and that correlations between models are not too high. The best situation is when diversification operates and outcomes of forking paths reveal complementary facets of the initial problem. These ideas have blossomed in the frequentist (Hansen (2007), Zhang (2015), Zhang and Liu (2019) and Zhu et al. (2023)) and Bayesian (Draper (1995), Raftery et al. (1997)) circles. For instance, Bayesian averaging has recently been used in Avramov et al. (2023) to cope with model uncertainty.

Nevertheless, we forcefully underline that model averaging and extreme bound analysis (EBA) are only special cases of forking paths. In most papers on model averaging and EBA, the data is fixed, and models are generated through alternative independent variable combinations. With forking paths, this is also allowed, but, more importantly, the ways to construct the initial sample can be discretionarily many, and estimators are not necessarily unique. Crucially, most results in model averaging consider that the number of models is finite, while the sample size increases to infinity. In the present paper, it is the opposite: sample sizes are arbitrarily small or large, and it is the number of paths that increases.

In addition to improving the robustness of reported results, framing empirical work as successive mappings helps organize code more neatly into a well-structured research pipeline. As is shown in Pérignon et al. (2023), coding skills are linked to the reproducibility of empirical studies. In our framework, each mapping has its own module, which opportunely prevents potential errors in lengthy scripts written in one block. It also forces to reflect upon the computational cost of each step of the research project and how to optimize it.[10] Consequently, an exhaustive approach

[9] See also de Prado (2023) for a similar categorization.

[10] Even if modern computers allow for the parallelization of tasks, the complexity of most pipelines out-

to the reporting of results compels the analyst to focus on the first order choices of the research process and to filter out the unnecessary artifices.

1.3 Summary of contributions

Beyond model averaging and inference, we propose several ways to operationalize forking paths. The first one seeks to determine which design choices have an impact on the distribution of outcomes. This is very important because if some modelling stages change the sign or significance of an effect, they must be transparently documented.

The second application pertains to *partial* replication, a task which we also call *confirmation*. The purpose here is not to exactly reproduce a given study but rather to test the generalization ability of a published result in slightly different contexts. To evaluate if a prior study has reported a *plausible* value, we compare it with the ones that we are able to produce with many paths. We devise an indicator, which we call the *Ease-to-Confirm* (EtC) and that measures the extent to which the original effect is compatible with the outcomes from the paths which we spanned. In essence, it measures the degree to which the original published value lies inside or outside the distribution of path-generated effects.

Lastly, a third use of paths relates to multiple testing (MT). Modern approaches to MT rely on bootstrapping. This assumes that the data available to the researcher (e.g., one path) is representative of the full distribution of the effect under scrutiny. Such an assumption may be excessively strong, which is why we propose to replace bootstrapped samples by forking paths, a method we refer to as **exhaustive multiple testing**. Because paths are more diverse, they are likely to produce more extreme outcomes and significance hurdles that are higher, compared to thresholds that originate from bootstrap-based MT techniques.

We illustrate our framework and operational recommendations with three empirical studies. The first one relies on the prediction of the equity premium, a well-documented research question in financial economics.[11] We consider a large number of ways to run the empirical protocol and report the distributions of test statistics. The latter allow us to determine which design choices alter the average of the statistics and are hence strong drivers thereof. For instance, switching from a simple OLS estimator to the Amihud and Hurvich (2004) specification has very little impact on coefficients. However, subsampling over two different periods generates substantial differences in estimates. In this first study, we also compare our results with those of Goyal et al. (2023) and find that their figures are mostly plausible and have magnitudes that can easily be reproduced.

Our second study revolves around asset pricing anomalies. Its core focus is on multiple testing whereby we seek to determine if the seemingly strongest factor is indeed strong enough. We compare the traditional approach to the one based on paths. We find that the improved thresholds generated by paths are significantly more conservative, compared to those of one traditional bootstrap-based method. From an investment standpoint, this means that our method implies that genuine anomalies are scarcer than the literature previously reported. In Harvey et al. (2016), the authors recommend to raise the significance threshold of t-statistics to 3.0. With standard bootstrap-based multiple testing, we recommend to push it to 4.5. If we rely on forking paths, the bar is set at least at 8.2, a level that few anomalies are able to pass. Unfortunately, a strong reduction of the risk of false positives comes with an increase of false negatives. In some situations, bad investments matter more than missed opportunities.

The requirement that a *genuine* factor must remain strong under many specifications shares some intuition with the literature on invariance-based causality (see Peters et al. (2016), Arjovsky et al. (2019) and Bühlmann (2020) for an overview). The premise therein is that in order to reveal a

weigh the CPU (and GPU/TPU) capabilities of standard machines. This is likely to limit the exploration of potentially promising but untested questions and configurations.

[11] Forking paths are tackled via the notion of *non-standard errors* for asset pricing anomalies in Soebhag et al. (2023) and Walter et al. (2023).

causal link between two variables, the relationship must hold in several sampling environments. If, for instance, each environment alters the distribution of X while preserving the either correlation between X and Y, or the conditional law of Y knowing X, then inferential conclusions are stronger than if they rely on one environment only. If the effect remains invariant across multiple studies or datasets (i.e., it is *replicable* outside its original sample), then it is robust, and possibly causal - this notion being out of the scope of the present paper.

Our third application revolves around Fama and MacBeth (1973) regressions and the estimation of risk premia. It reveals that while some choices do not matter much (winsorizing loadings after the first pass), others are more critical, especially the sampling of returns before the first pass. We compare our results with those of Fama and MacBeth (1973) and Ang et al. (2020). We find that for the former, the reported figures are entirely reliable. For the latter, some market premia are realistic, but others are not.

For each one of these three topics in asset pricing, we exhibit our *Ease-to-Confirm* metric, based on published results from the literature. We document heterogeneity in the outomes: while some prior results can easily be corroborated, others fall out of the range of the paths which we ware able to span. This leads us to conclude that some findings may be considered as robust and replicable, while other may be attributable possibly to luck or data mining (see also Chen et al. (2023)).

The remainder of the paper is structured as follows. Section 2 lays out a representation of research studies as compositions of operators. Therein, we also propose several ways to exploit the paths generated by these compositions, for instance by characterizing which mappings are significant for a given research output, or by means of model averaging. An important benefit from paths is that they allow to locate prior research results within the distribution from the paths: this allows to corroborate these prior findings - or not. Our ideas are illustrated through three empirical studies which are mentioned throughout the paper and are located in Sections 3, 4 and 5. Finally, Section 6 concludes. The appendix features some supplementary material.

2 Theoretical groundwork

This section comprises the analytical background of the paper which models the research process as compositions of operators. The foundations are laid in subsections 2.1 and 2.2. The methods for the operationalization of the paths are presented in the remaining four subsections.

2.1 Overarching framework and p-hacking

2.1.1 Notations

We start with a few conventions on notation. Unless otherwise stated, the integer N will always be the length of the vectors and the number of rows of matrices. Henceforth, lowercase bold letters $\mathbb{d} = \{\mathbb{d}_1, \dots, \mathbb{d}_N\}$ will denote vectors and uppercase bold letters matrices or tables. For the latter, we adopt the **tidy data** convention of Wickham (2014): rows are observations and columns are variables. Finally, we will sometimes (when there is little ambiguity) use the simplified notation $f(\mathbb{d})$ for the vector $[f(\mathbb{d}_1), \dots f(\mathbb{d}_N)]$. Moreover, for a matrix $\boldsymbol{M}$ or a vector $\boldsymbol{v}$, $\boldsymbol{M}'$ and $\boldsymbol{v}'$ will denote their transpose.

In addition, we will often compare two alternative inputs. Readers are accustomed to $\boldsymbol{X}$ and $\boldsymbol{y}$ for modelling purposes. To avoid any confusion, we work with the letters $\mathbb{D}$ and $\boldsymbol{D}$, which will stand for two versions of some data that is collected and then possibly transformed by the researcher. This choice of notation is disconcerting at first, but imperative because we will restrict the use of $\boldsymbol{X}$ and $\boldsymbol{y}$ letters to linear models later on.

We assume that the empirical part of research process starts with some input which we call $\mathbb{D}$ and can be thought of as the initial version of the data that is collected. The study is modelled as

a sequence of operations f_j that occur successively so that the reference research output o_J (e.g., one t-statistic) is such that

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

where $f_j : S_j \mapsto S_{j+1}$, with S_1 and S_{J+1} encompassing the sets of feasible input $\mathbb{D}$ and output values, respectively. For simplicity, we can assume that o_J is simply a real number, but it may be a more complex object, such as a vector (e.g., confidence interval) or a matrix. The index J indicates that the output is the result of J successive operations, which Gelman and Loken (2014), among others, refer to as *forking paths*. Examples of such operations are provided in Appendix A and include for instance: missing data imputation or removal, winsorization, variable selection, variable scaling, subsampling, choice of estimator, etc. Here, the output depends on J and the sequence of mappings f_j. Later on, for simplicity, we will index outputs with p, which will be the index of the corresponding path.

More precisely, we can write the output-generating process as

$$o_J(\mathbb{D}, \boldsymbol{P}) = \left[\bigcirc_{j=J}^{1} f_{j,\boldsymbol{p}_j}\right](\mathbb{D}) = f_{J,\boldsymbol{p}_J} \circ \cdots \circ f_{1,\boldsymbol{p}_1}(\mathbb{D}), \tag{2}$$

where $\boldsymbol{P}$ encompasses the parameter sets $\boldsymbol{p}_j$ for all the mappings. These parameters may be fixed, or random, e.g., when sampling arbitrary thresholds for winsorization. The output o_J is a random variable that depends on the realizations of the operators f_j - and possibly on that of $\mathbb{D}$ if stochastic initial samples are allowed. In all generality, the realization of f_j may depend on those of prior operators (f_i for $i < j$). Plainly, the order of mappings may matter: it does not make sense to perform data imputation *after* estimation.

Henceforth, we assume that any mapping f_j has r_j deterministic options which the researcher must choose from, and which we write $\mathbb{f}_{j,r}$, for $r = 1, \dots r_j$, where $r_j \geq 2$. For example, this can be alternative ways of handling missing data (deletion versus imputation), or the set of possible combinations of independent variables, in which case r_j is the cardinal of this set (i.e., all permutations that are relevant for the study). This makes $P = \prod_{j=1}^{J} r_j$ paths in total.

Paths are determined by the choice of their options for each layer. Thus, we define paths as $p := \{\mathbb{f}_{j,r_{p,j}}\}_{1 \leq j \leq J}$, where $r_{p,j}$ is the option choice of path p for layer j. To ease notation, we will sometimes write $\mathbb{f}_{j,r(p)}$ for the option of layer j though which path p passes.

A crucial facet of forking paths is their diversity, which we can measure via their proximity, or lack thereof. For each layer j, we assume we can define a distance function d_j between all choices of the layer, e.g., between paths p and q: $d_j(p, q) = d_j(\mathbb{f}_{j,r(p)}, \mathbb{f}_{j,r(q)})$. We can then aggregate into a total distance between two paths

$$d(p, q) = \sum_{j=1}^{J} \omega_j d_j(\mathbb{f}_{j,r(p)}, \mathbb{f}_{j,r(q)}), \tag{3}$$

where ω_j specifies the relative importance of layer j. A simple and explicit form for $d(p, q)$ will be used in Section 2.2.

2.1.2 Divergence of outputs

One interesting question pertains to the sensitivity of the output o_J to a change in initial input $\mathbb{D}$. In order to derive theoretical results, we must impose some conditions on the mappings f_j and we choose to work with Lipschitz smoothness. More precisely, we assume that for $\mathbb{D}, \boldsymbol{D} \in S_j$, there exists some constant $c_j > 0$ such that

$$\|f_j(\mathbb{D}) - f_j(\boldsymbol{D})\| \leq c_j \|\mathbb{D} - \boldsymbol{D}\|, \tag{4}$$

for some norms which are implicitly defined on S_{j+1} and S_j. For simplicity, we do not precisely define these norms, except in Appendix A in which we explicit some Lipschitz constants for a

few specific operators. In all generality, the object $\mathbb{D}$ can comprise several data types, categorical features notably (ordinal or nominal). Handling distances with such features is complex, though not impossible (see Boriah et al. (2008)), but for the sake of simplicity, the exposé will essentially assume that $\mathbb{D}$ and $\boldsymbol{D}$ are matrices of real numbers.

Composing two operators yields

$$\begin{aligned}\|f_{j+1} \circ f_j(\mathbb{D}) - f_{j+1} \circ f_j(\boldsymbol{D})\| &\leq c_{j+1}\|f_j(\mathbb{D}) - f_j(\boldsymbol{D}\| \\ &\leq c_j c_{j+1}\|\mathbb{D} - \boldsymbol{D}\|.\end{aligned}$$

Iterating this inequality leads to

$$\left\|\left[\bigcirc_{j=J}^{1} f_j\right](\mathbb{D}) - \left[\bigcirc_{j=J}^{1} f_j\right](\boldsymbol{D})\right\| \leq \|\mathbb{D} - \boldsymbol{D}\| \prod_{j=1}^{J} c_j. \tag{5}$$

More generally, the accumulation of shifts may not start at the initial data sample $\mathbb{D}$, but at a later stage, say at o_K, after K steps, for $K < J$. It is easy to prove the following lemma.

Lemma 1. *If o_J is given by Equation (1) and the mappings f_j satisfy (4), then for $1 \leq K < J$,*

$$\left\|\left[\bigcirc_{j=J}^{K+1} f_j\right](o_K(\mathbb{D})) - \left[\bigcirc_{j=J}^{K+1} f_j\right](o_K(\boldsymbol{D}))\right\| \leq \|\mathbb{D} - \boldsymbol{D}\| \prod_{j=K+1}^{J} c_j. \tag{6}$$

It is obvious that the constants c_j are the main drivers of the error bounds. In practice, a lower bound for the c_j is often 1, meaning that their compounded effect can be sizeable, theoretically, especially if many c_j are such that $c_j \gg 1$. Thus, adding steps in the design is likely to increase the amplitude of the difference between outcomes. If the latter are scalars and the norm is the max-norm, then we recover the width of the so-called *p-hacking interval* introduced in Coker et al. (2021).

For illustration purposes, let us consider the sample mean, which is abundantly used, if only in summary statistics. If f is the **sample mean** operation, we have, via Hölder's inequality in the last inequality,

$$\|f(\mathbb{d}) - f(\boldsymbol{d})\|_p = \left|\frac{1}{N}\sum_{n=1}^{N} \mathbb{d}_n - \frac{1}{N}\sum_{n=1}^{N} d_n\right| \leq \frac{1}{N}\sum_{n=1}^{N} |\mathbb{d}_n - d_n| \leq N^{-1/p}\|\mathbb{d} - \boldsymbol{d}\|_p, \tag{7}$$

so that in this case the Lipschitz constant is $N^{-1/p}$.

We list other important mappings along with upper bounds on their Lipschitz constants in Appendix A. Few are exactly sharp and in some cases, they even depend on the inputs, $\mathbb{d}$ and $\boldsymbol{d}$, or $\mathbb{D}$ and $\boldsymbol{D}$.

The upper bound for the range of outcomes provided in Lemma 1 is theoretical. Indeed, it is possible that conflicting effects in the mappings yield actual outcomes that are less dispersed. It is only after the paths have been spanned that we can evaluate the divergence, as in Menkveld et al. (2023) for instance. This issue is postponed to Section 2.3.

2.1.3 p-hacking

In practice, the paths are not necessarily spanned meticulously. Data snooping is a powerful generator of paths because it requires to reflect on the methodological choices underpinning the initial baseline. Let us now assume that paths and outcomes $o^{(p)}$ are indexed by some integer p which represents the order in which the spanning has occurred: $o^{(1)}$ is the first output, etc.

Then, **simple *p*-hacking** is simply the process of generating a sequence $o^{(p)}$ and stopping when the p^{th} output is deemed satisfactory: it will be the one that is retained for publication. This usually requires two conditions. The first one is statistical significance, i.e., the outcome must be associated with a confidence level $1-\alpha$ that is high enough, usually 95% or 99%. The second condition is the sign of the outcome. Most of the time, the direction of the effect is important and it may occur that preliminary results be significant, but in the wrong direction. In which case, more paths need to be explored, unless the original hypothesis is validated, or revised, as in HARKing (Kerr (1998), Hollenbeck and Wright (2017)). Note however that if $o^{(1)}$ is already statistically significant and in the desired direction, then p-hacking does not need to start. **Robust *p*-hacking** is more stringent. Here we do not require that only one path be fitting a narrative, but a neighborhood of paths. Thanks to this, it is possible to present deviations from the default path (robustness checks) that corroborate and strengthen its conclusions.

Finally, we mention the ultimate level of snooping, which we call **vicious *p*-hacking**. This refers to the case when paths are explored until a sufficient amount of p-values can be extracted in such a way that they successfully pass a test that detects p-hacking, for instance those mentioned in Elliott et al. (2022). Because such tests are very recent, it is highly unlikely that vicious p-hacking has already occurred. Nevertheless, this notion is interesting because, given a vector of p-values generated by forking paths, we can try to evaluate the potential for vicious p-hacking.

Let us be more specific. Suppose we focus on the common case of one-sided t-tests, which we will cover in Section 4 for anomaly detection. In Elliott et al. (2022), it is shown in Theorem 2 that, under mild technical assumptions and in the absence of p-hacking, the density of the p-values should be decreasing and convex. Because of their importance, discuss the relevance and applicability of p-hacking tests for forking paths in Appendix C.

Let us assume that we have spanned many paths and obtained the corresponding p-values, which are distributed on the unit interval. We split the first half of this interval into I sub-intervals of equal sizes and we write n_i for the number of p-values inside these intervals. Simply put, the n_i are simply the count values of the histogram.

Applied to sample values, the first condition for the absence of p-hacking is that $n_i > n_{i+1}$ for $i = 1, \dots, \lfloor (I-1)/2 \rfloor$. If it is satisfied, the second order condition requires that $n_i/n_{i+1} > n_{i+1}/n_{i+2}$. Given the actual values of n_i, it is possible to evaluate the number of these inequalities that are fulfilled. If some of them are not, we can also assess the extent to which they are violated. This yields a qualitative judgement of how much trafficking in the distribution is needed before it can pass a p-hacking detection test. This will be illustrated in Section 4.5.

2.2 Paths as pseudo-environments

2.2.1 Invariant effects

In a large number of studies, the aim is to determine if a particular effect holds. For simplicity, let us assume that such effect can be summarized as a scalar, b, and that the null is $b = 0$, so that the researcher seeks to reject it in order to obtain a publishable result. Importantly, the effect may not be constant and, in all generality, it can be viewed as a random variable that depends on the environment through which it is evaluated. This line of reasoning is in fact pregnant in several fields:

- For instance, in causal inference, it has been proven that if an effect remains invariant across several sampling environments, then this effect can be considered as causal in some sense (Peters et al. (2016), Pfister et al. (2019)). Causality is out of the scope of the present paper, but the analogy between paths and environments is at the core of our reasoning.
- Likewise, the idea that multiple models are useful to characterize variable importance is also popular in machine learning (Fisher et al. (2019)).

- Lastly, a final parallel can be drawn with the social sciences, for which the *observer effect* may alter the gathering of data: the characteristic of a researcher can influence the behavior of the subjects being studied (Monahan and Fisher (2010)).

One important premise of the paper is that effects are indeed random, and that, most of the time, researchers seek to characterize the *average effect*, $\mathbb{E}[b]$. For instance, in finance, it is well-known that risk premia depend on the economic environment. Gagliardini et al. (2016) have shown that traditional factor premia are strongly dependent on NBER cycles. Hence, premia will depend on the time-frame on which they are evaluated, and, more generally, on the data and estimator that is used. Consequently, to have a better understanding thereof, it is valuable to generate not only a representative average, but many plausible values that will serve to determine the distributions of the premia, taken as random variables.

Said differently, there is considerable value in capturing effects in several environments. For instance, if estimated effects $\hat{b}_p$ are positive (and possibly statistically significant) across P environments, there is arguably more evidence in favor of it than from a single point estimate. This is exactly the rationale of Peters et al. (2016), a seminal contribution on invariance-based inference. Therein, the authors argue that a set of causal predictors (parents in a directed acyclic graph) can be obtained by taking the intersection of variables which are significant across all available environments. In the present paper, forking paths replace environments, and we refrain from invoking causal effects. Paths are simply used to shed light on multitudes of facets of effects and we exploit these facets to provide a more exhaustive characterization of effects.

2.2.2 Stylized conditions for convergence

Henceforth, we *want* (and sometimes need) to assume that estimated effects $\hat{b}_p$ are such that their distribution converges to the true distribution of b, as the number of paths, P, increases to infinity. Below, we outline conditions under which this may occur. Fundamentally, we are seeking a concentration inequality and such types of results are well-known when the underlying variables are i.i.d. The major issue here is that it is not realistic to assume that outcomes from paths are independent variables. Recently, there have been advances in concentration inequalities for correlated sequences.[12] However, the central result we rely upon is Theorem 1 of Azriel and Schwartzman (2015), which we recall below.

One way of representing our line of thought is to write the sought effect as $b = \bar{b} + \tilde{b}$, i.e., the sum of a constant $\bar{b}$, the mean effect, plus a zero mean Gaussian random term $\tilde{b}$ with variance σ_b^2. For the remainder of the section, we thus propose the following generic data generating process

$$Y_p = X_p(\bar{b} + \alpha_p \tilde{b}) + (1 - \alpha_p)\epsilon_p, \quad p = 1, \dots, P, \tag{8}$$

where p indexes the path (environment) and α_p is a deterministic perturbation intensity, which, in all generality can be path-specific. The error terms ϵ_p are assumed to be independent from $\tilde{b}$ and exogenous, i.e., such that $\mathbb{E}[\epsilon_p | X_p] = 0$. All vectors Y_p, X_p and ϵ_p have the same length. However, from one path to another, this length may vary, as different environments may entail changing sample sizes. For a given path p, the standard OLS estimator is then

$$\hat{b}_p = (X_p' X_p)^{-1} X_p' Y_p = \bar{b} + \alpha_p \tilde{b} + (1 - \alpha_p) e_p, \quad e_p = (X_p' X_p)^{-1} X_p' \epsilon_p. \tag{9}$$

The shrinkage intensity $\alpha_p \in [0, 1]$ implies that the estimator does not capture the entirety of the random component of b. Plainly, e_p is a blurring term that depends on path p and the shrinkage form is commonplace in contamination models (e.g., from Huber (1964) to Chen et al. (2016)), wherein the data imperfectly reflects reality. In short, the estimated effects capture the average, but

[12] We refer to either of Berry-Esseen results (Bentkus et al. (1997), Jirak (2016, 2023)), or to Dvoretzky–Kiefer–Wolfowitz inequalities (Kontorovich and Weiss (2014), Chen and Wu (2018)).

they are perturbed by some independent Gaussian noise. Note that if the perturbation has variance exactly equal to $\sigma_e^2 = \frac{1+\alpha_p}{1-\alpha_p}\sigma_b^2$, then $\sigma_b^2 = \sigma_{\hat{b}_p}^2$ and the estimator $\hat{b}_p$ has the same distribution as b. This *strong* assumption is required if we need L^2 convergence of the full distribution of the empirical $\hat{b}_p$. If we are only interested in the *average* effect, the variance term matters much less, as long as it is finite. Finally, the dependence between two estimators from two different paths can be shown to be

$$\mathbb{C}\text{or}(\hat{b}_p, \hat{b}_q) = \alpha_p\alpha_q + (1-\alpha_p)(1-\alpha_q)\mathbb{C}\text{or}(e_p, e_q), \tag{10}$$

which depends on the shrinkage intensities and on the correlation between the paths' perturbations.

To theoretically show the benefits of forking paths, we then rely on a stylized special case of the above framework. We assume for simplicity and tractability that estimated effects across all paths $\hat{b}_1, \ldots, \hat{b}_P$ follow standard $N(0,1)$ Gaussian variables with correlation matrix $\mathbf{\Sigma}_P$ that gathers all pairs of correlations defined in Equation (10). We underline that general Gaussian variates can also be considered since they can be normalized (demeaned and scaled). Then, from Azriel and Schwartzman (2015), there exists a constant $c > 0$ such that

$$\sup_{x\in\mathbb{R}} \mathbb{E}\left[(\Phi(x) - \Phi_{b,P}(x))^2\right] \leq \frac{1}{4P} + c\|\mathbf{\Sigma}_P\|_1, \tag{11}$$

where $\Phi(x)$ is the Gaussian cdf of the true effect and $\Phi_{\hat{b},P}(x)$ the empirical distribution of estimated effects $\hat{b}_p$, obtained by the generation of P paths. We thus seek conditions under which $\|\mathbf{\Sigma}_P\|_1$ will shrink to zero as $P \to \infty$.

Therefore, it remains to characterize the correlation between path outcomes. As advocated earlier, it makes sense that it be driven by the proximity between paths. Hence we postulate that the correlations between paths p and q are such that $\rho(p,q) = \mathbb{C}\text{or}(\hat{b}_p, \hat{b}_q) = h(d(p,q))$, where $d(\cdot,\cdot)$ is the distance function defined in Equation (3) and h is some strictly decreasing function from $\mathbb{R}_+$ to $[-1,1]$ such that $h(0) = 1$. Because of Equation (11), we are interested in

$$\|\mathbf{\Sigma}_P\|_1 = P^{-2}\sum_{1\leq p,q\leq P}|\rho(p,q)| = P^{-1} + 2P^{-2}\sum_{1\leq p<q\leq P}|h(d(p,q))|. \tag{12}$$

Without further assumptions, it is impossible to provide additional information on the empirical distribution of the distances. We are bound to specify a more explicit form for $h \circ d$. To do so, we first posit a simple distance function which corresponds to the number of layer choices that differ between p and q:

$$d(p,q) = \#\{j, r_{p,j} \neq r_{q,j}\} \in \{0, 1, \ldots, J\}, \tag{13}$$

where the operator $\#\{A\}$ measures the cardinal of set A and we recall that $r_{p,j}$ is the mapping option through with path p passes for layer j. Because there are J layers in the protocol, we consider that the number of non-identical mapping options can be a good proxy for the differences between the related outcomes. If two paths have only one non-common choice, then, their outcomes should be more correlated than if they have zero commonality. As we show in Appendix B, this simplification allows to fully characterize the distribution of distances and to determine conditions for convergence as summarized in the following result.

Proposition 2. *Assuming* $\rho(p,q) = \rho^{d(p,q)}$ *with* d *given by* (13)*, we have that* $\|\mathbf{\Sigma}_P\|_1 \overset{J\to\infty}{\to} 0$*, and, by* (11)*,* $\Phi_{\hat{b},P}$ *converges uniformly to* Φ *in* L^2*.*

One important feature of the proposition is that the number of paths has to grow to infinity via the number of mappings J, not via the number of mapping options r_j. As we show in the proof in Appendix B, if J remains finite, the norm does not decrease to zero and convergence to the true cdf does not occur. In practice of course, J will be finite but should be large, so that the norm of the matrix will be small, thereby implying small errors on the true cdf.

Proposition 2 remains a theoretical result, as it requires ways to measure or estimate $\rho(p,q)$. Indeed, in our framework, we are endowed with only one outcome for each path, which makes correlation evaluation impossible. As Burnham and Anderson (2004a) put it bluntly: "*have no basis to estimate the across-model correlation*". One way to bypass this hurdle would be to resort to a resampling of the original datasets that serve as input to the paths, $\mathbb{D}$. This important topic is investigated in Section 4.6 wherein we confirm the main assumption of Proposition 2.

Another route toward convergence would be to assume that beyond a certain distance (e.g., $d(p,q) > 7$), two paths are sufficiently dissimilar so that we can assume a zero correlation, or even independence. This relates to so-called m-dependent variables and their link to central limit theorems is investigated in Hoeffding and Robbins (1948), Diananda (1955) and Orey (1958). But this is not the road we follow here.

2.3 The range of outcomes and its rate of increase

Another essential question which arises from paths, and underlined in Lemma 1, is the speed at which extreme outcomes diverge as a function of J, the number of mappings (i.e., design choices). Practically, it is impossible to assess the impact of each layer on the dispersion of, say, t-statistics, because all mappings are chained and outcomes are only produced by the final layer.

We thus propose a tractable method to evaluate the impact of the richness of the protocol (the number of paths) on the range of outcomes. First, we set an integer $K \in [1, J-1]$ which will correspond to the number of mappings that are fixed. For each K, we write $c_l(K)$ for each combination (set) of fixed mappings; they are indexed by $l = 1, \ldots, \binom{J}{K}$. Each $c_l(K)$ is associated with $\prod_{k=1}^{K} r_{k \in c_l(K)}$ fixed configurations and each configuration has $\prod_{k=1}^{J-K} r_{k \notin c_l(K)}$ possible paths. The rationale is that by fixing K mappings, we leave $J-K$ as degrees of freedom in the protocol, as if there were in fact $J-K$ actual layers of design choices. Furthermore, testing all permutations allows to obtain a very rich collection of cases.

One such case is illustrated in Figure 1 for $J = 4$ and $K = 2$. There are $\binom{4}{2} = 6$ possible combinations of 2 mappings and we show one of them with the grey circles: we fix the first and third layers. Given the number of options, there are then 4 combinations for the fixed mappings, and 9 paths can be followed for these 4 combinations (when the first and third mappings are fixed to one of their possible alternatives).

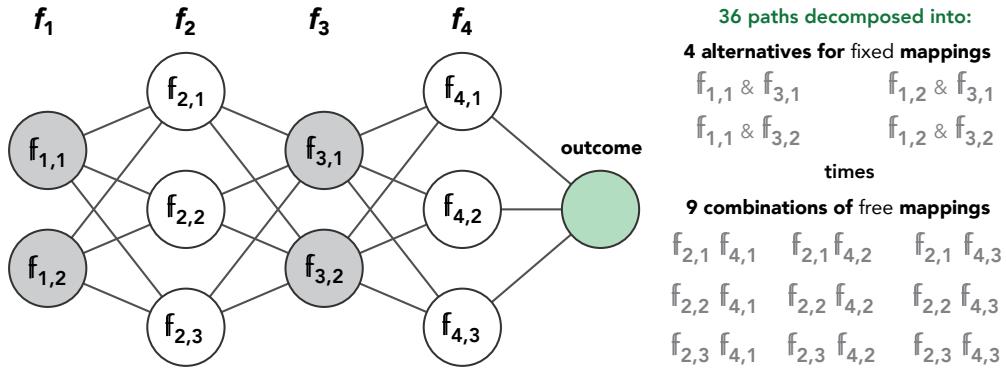

Figure 1: **Spanning the paths: fixed versus free mappings**. The above representation features $J = 4$ mappings with a number of options $r_1 = r_3 = 2$ and $r_2 = r_4 = 3$. The grey circles show the ($K = 2$) **fixed mappings** while the white ones pertain to the **free** mappings.

For each $c_l(K)$ (i.e., one configuration like the one shown in Figure 1), it is possible to compute some statistics over all the outcomes generated by the free mappings. We write $S_{K,l}$ for the set of

paths related to $c_l(K)$ and we are interested in the breadth of the outcomes related to the paths, which, following Coker et al. (2021), we call the range of *hacking intervals*:

$$I_{K,l} = \max_{p \in S_{K,l}} o_p - \min_{p \in S_{K,l}} o_p. \tag{14}$$

Given these outputs, it is possible to calculate aggregate values, notably the average range of intervals, for each K:

$$ARI(K) = \frac{1}{n_K} \sum_{l=1}^{n_K} I_{K,l}, \tag{15}$$

where $n_K = \sum_{l=1}^{\binom{J}{K}} \prod_{k=1}^{K} r_{k \in c_l(K)}$ is the total number of intervals for a fixed K. As K increases, the number of free mapping decreases, and thus, the ARI is expected to decrease. In the spirit of Lipschitz constants mentioned in the previous section, we define the **growth rate of average intervals**, as the number of free mappings (i.e., $J - K$) increases:

$$\rho_{J-k} = \frac{ARI(J-k)}{ARI(J-k-1)} - 1. \tag{16}$$

This rate is not supposed to be uniform with k. Indeed, in the early stages, the intervals are likely to expand rapidly, but, as $J - k$ increases, ρ_{J-k} should shrink towards zero, so that additional design choices only impact the range of outcomes marginally. This will be confirmed in one of our empirical study in Section 3.6 below.

2.4 Inference from paths

Naturally, once the paths have been generated comes the question of the exploration of their outcomes. In this section, we revert back to the ubiquitous need for inference in empirical research when conclusions mostly depend on the statistical significance of a given effect. Below, we consider the question of the confidence intervals for the *average* effect in Section 2.4.1. To do so, we exploit a few ideas from the literature on model averaging. We refer to Moral-Benito (2015), Zhang and Liu (2019) and Steel (2020) for surveys on the matter. In Section 2.4.2, we outline our definition of *conditional* averages.

2.4.1 Model averaging

The expression for the estimated average effect is always written as

$$\hat{b}_* = \sum_{p=1}^{P} w_p \hat{b}_p, \quad \text{with} \quad \sum_{p=1}^{P} w_p = 1, \tag{17}$$

where the main issue is the determination of the weights w_p. This stage is very far from benign, as we will show and highlight in our empirical applications. When weights deviate from uniformity ($w_p = P^{-1}$), it means that paths are far from equal in generating trustworthy values for effects.

We first consider a *frequentist* average of the effects and follow Buckland et al. (1997) and Burnham and Anderson (2004a). We define positive weights that rely on likelihood through information criteria:

$$w_p = \frac{e^{-\Delta_p/2}}{\sum_{k=1}^{P} e^{-\Delta_k/2}}, \quad \Delta_p = AIC_p - \min_p AIC_p, \tag{18}$$

where AIC_p is the Akaike Information Criterion of model (i.e., path) p. For the estimation of the variance of the aggregate estimator, we follow Equation (1) in Burnham and Anderson (2004b)

who make the conservative assumption of perfect correlation between estimators ($\mathbb{C}\text{or}(\hat{b}_p, \hat{b}_q) = 1$):

$$\hat{\sigma}_*^2 = \left(\sum_{p=1}^{P} w_p \sqrt{\hat{\sigma}_p^2 + (\hat{b}_* - \hat{b}_p)^2} \right)^2 . \tag{19}$$

Assuming mild dependence conditions in order to be able to invoke the Central Limit Theorem to derive confidence intervals for $\bar{b}$ at the α-level, we have:

$$\mathbb{P}\left[\bar{b} \in \left(\hat{b}_* - c_{\alpha/2}\hat{\sigma}_*/\sqrt{P},\ \hat{b}_* + c_{\alpha/2}\hat{\sigma}_*/\sqrt{P} \right) \right] = 1 - \alpha, \tag{20}$$

where c_α is the quantile function of the standard normal law. Importantly, note that the speed of convergence, in contrast to most of the literature on model averaging for linear models,[13] is not in the sample size, but in the **number of paths**.

For the sake of completeness, we also propose a weighting scheme from the perspective of **Bayesian** model averaging. We follow the standard nomenclature, as is for instance laid out in Hoeting et al. (1999). The quantity of interest is b, with posterior probability given the data D equal to

$$\mathbb{P}[b|D] = \sum_{p=1}^{P} \mathbb{P}[b|M_p, D]\mathbb{P}[M_p|D],$$

where $\mathbb{M} = \{M_p, p = 1, \dots, P\}$ is the set of models under consideration. In this paper, one model corresponds to one complete path. Notably, the above equation translates to the following conditional average and variance:

$$\mathbb{E}[b|D] = \sum_{p=1}^{P} \hat{b}_p \mathbb{P}[M_p|D] \tag{21}$$

$$\mathbb{V}[b|D] = \sum_{p=1}^{P} \left(\mathbb{V}(b|M_p, D) + \hat{b}_p^2 \right) \mathbb{P}[M_p|D] - (\mathbb{E}[b|D])^2 \tag{22}$$

where $\hat{b}_p$ is the estimated effect from path p. The posterior model probabilities are given by

$$\mathbb{P}[M_p|D] = \left(\sum_{j=1}^{P} \frac{\mathbb{P}[M_j]}{\mathbb{P}[M_p]} \frac{l_D(M_j)}{l_D(M_p)} \right)^{-1} , \tag{23}$$

with $l_D(M_j)$ being the marginal likelihood of model j. Because we are agnostic with respect to the relative importance of paths, we set the prior odds $\frac{\mathbb{P}[M_j]}{\mathbb{P}[M_p]}$ equal to one. Note that in this case, the posterior probabilities are then simply proportional to the likelihoods. The remaining Bayes factor is by far the most complex and we follow the recommendations of Fernandez et al. (2001) (Equation (2.16), adapted for inhomogeneous sample sizes):

$$\frac{l_D(M_j)}{l_D(M_p)} = \left(\frac{n_j}{n_j + 1} \right)^{\frac{k_j}{2}} \left(\frac{n_p + 1}{n_p} \right)^{\frac{k_p}{2}} \frac{\left(\frac{s_p + n_m v_p}{n_p + 1} \right)^{(n_p - 1)/2}}{\left(\frac{s_j + n_j v_j}{n_j + 1} \right)^{(n_j - 1)/2}}, \tag{24}$$

where n_j is the inverse of the number of observations used in model j and k_j is the number of predictors in this model, omitting the constant ($k_j = 1$ in our case). Moreover, $n_j v_j$ is the sample variance of the dependent variable in model j. Finally, s_j is the sum of squared residuals under model j.

Averages of the form (17) along with confidence intervals (20) will be shown in Subsection 3.3 both with frequentist (18) and Bayesian (23) weights.

[13] We point for instance to Zhang and Liu (2019) and the references therein.

2.4.2 Conditional averages

Arguably one of the most important question with forking paths pertains to the impact of choices that researchers make on the distribution of the outcomes they generate. In Equation (2), the randomness in outcomes stems from the original sample $\mathbb{D}$, but also, and more importantly, from all modelling steps f_j that constitute each path. One interesting extension pertains to the random variables $\hat{b}_p|\{f_j = \mathbb{f}\}$, which are the values of effects, conditional on the knowledge of one layer choice for mapping j, say, $\{f_j = \mathbb{f}\}$. Therefore, $\hat{b}_p|\{f_j = \mathbb{f}\}$ gathers all realizations of $\hat{b}_p$ such that the path p "*passes*" through the layer option $\{f_j = \mathbb{f}\}$. This notation will serve to determine if the operator f_j has an important impact on the distribution of the outcome.

In order to test if one operator (i.e., layer) has an impact on the estimated outcomes, we define

$$\hat{b}_p^{\{f_j=\mathbb{f}\}} \quad \text{and} \quad \hat{b}_{-p}^{\{f_j=f\}} \tag{25}$$

as the random variables $\hat{b}_p|\{f_j = \mathbb{f}\}$ and $\hat{b}_{-p}|\{f_j = f\}$, respectively, for two different mapping alternatives $\mathbb{f} \neq f$ at layer j. Note that we introduce a special notation with a negative index. The variable $\hat{b}_{-p}^{\{f\}}$ corresponds to the outcome of the path which is exactly the same as path p, except for layer j. In other words, the two effects defined in (25) come from paths which are very close and have only one difference. This is depicted in Figure 2 below. The test layer (f_j) has two options (e.g., winsorizing or not) and we discriminate the paths based on whether they pass through one option (in **dotted blue**) or the other (in **orange**).

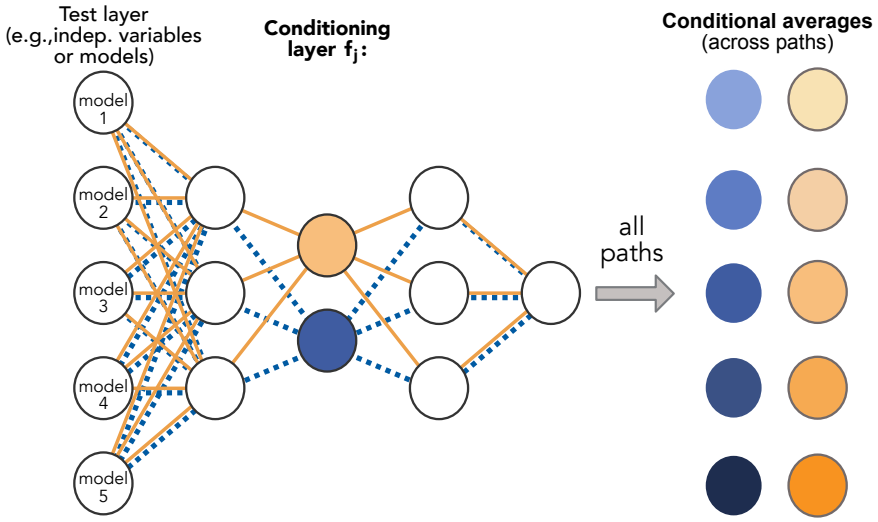

Figure 2: **Spanning the paths: passing through alternative mapping options**. We depict the notion of average effects, conditional on some layer choices. For instance, the variables $\hat{b}_p^{\{f_j=\mathbb{f}\}}$ corresponds to the outcomes of all dotted **blue paths** while $\hat{b}_{-p}^{\{f_j=f\}}$ to the outcomes of all **orange paths**.

Once the set of paths has been separated in two (or more, depending on the number of layer options), it is possible to test whether the distributions of $\hat{b}_p^{\{\mathbb{f}\}}$ and $\hat{b}_q^{\{f\}}$ are different. While tests on the full distribution (Kolmogorov-Smirnov) are available, we prefer to stick to simple mean tests. However, because estimation outputs are weighted prior to inference, we need to apply the corresponding weights before averaging (w_p and w_{-p}). Thus, we can resort to a simple Student t-test to the series of outputs

$$\Delta_p = \frac{2}{P}\left(w_p \hat{b}_p^{\{f_j=\mathbb{f}\}} - w_{-p}\hat{b}_{-p}^{\{f_j=f\}}\right), \tag{26}$$

where the $2/P$ scaling is there to ensure that a simple mean of the above series over the $P/2$ paths will be exactly equal to the difference in weighted averages. This can easily be generalized to decision layers with more than two options.

One canonical option for f_j is the subsample. For instance, Goyal et al. (2023) report their results for the first and second halves of their samples. Their results underline the marked sensitivity of reported effects to the choice of the period. We will forcefully corroborate this phenomenon. The application of the above ideas are located in all of our empirical analyses, in Subsections 3.4, 4.4, and 5.3.

2.5 Exhaustive multiple testing

Not all mappings are equal. For instance, the choice of independent variable may very well be a strong modelling assumption. Consider the two alternative questions:

- Does variable X predict variable Y?
- Can variable Y be predicted?

In the first case, it the focus is clearly on the predictive ability of variable X, hence picking it as independent variable is not a design choice, it is an imperative. In the second option, choosing variable X or Z, or W is left to the appreciation of the researcher, and, in fact, it is conceivable to mine as much data as possible to find the few predictors that may indeed predict Y. Typically, in the debate on the predictability of the equity premium, dozens of variables have been proposed. If many are studied, statistical significance must be corrected for **multiple testing**.[14] Recently, several studies in finance have approached research questions by resorting to large scale tests in which many predictors are considered (Yan and Zheng (2017), Chordia et al. (2020), Giglio et al. (2021) and Jensen et al. (2021)).

The approach we have advocated until now is slightly different. We have assumed that the researcher has a precise research question in mind, but that there are many different ways to answer it, thanks to small shifts, or tweaks, in the empirical protocol, exactly as in Huntington-Klein et al. (2021) and Menkveld et al. (2023). The difference with multiple testing is illustrated in Figure 3. In the left graph, the space of models is wide, and all hypotheses are tested with the same unique protocol (e.g., simple portfolio sorts, or linear models). In the right plot, the scope is narrower, and the number of hypothesis is small (e.g., just one), but the methods used to reach conclusions are heterogeneous. In financial economics, the first type can be found in Jensen et al. (2021), in which the authors approach the topic of asset pricing anomalies via a large-scale study encompassing thousands of firms worldwide and testing hundreds of factors. The portfolios are all constructed using the same methodology. Two examples of the second type of studies are Asness and Frazzini (2013) and Amenc et al. (2020), wherein the authors focus solely on the **value** anomaly, but propose alternative ways to construct value factors. Similar analyses have been carried out for the **momentum** anomaly (see Novy-Marx (2012) and Gong et al. (2015)), and, more generally, to a broad scope of anomalies in Soebhag et al. (2023) and Walter et al. (2023).

In this subsection, we argue that it is in fact possible to combine the two, i.e., to enhance multiple testing with exhaustive protocols. The rationale is the following. Most state-of-the-art techniques used in multiple testing (see Harvey et al. (2020)) rely on bootstrapping. Now, as is shown in Romano and Wolf (2005) (Assumption 3.1), this requires that the sampling distribution

[14] Because we propose to generate series of outcomes, the ideas presented in the present paper are undoubtedly linked to this notion, a theme that goes back at least to Bonferroni (1936), and which is applied in serveral disciplines, including medicine (Farcomeni (2008)), economics (Viviano et al. (2022)), finance (Harvey and Liu (2020), Harvey et al. (2020), Giglio et al. (2021)), generic model discrimination (Hansen et al. (2011)) and statistics more generally (Fan and Han (2017), Wang et al. (2017) to cite but a few). In many cases, as in Romano and Wolf (2005, 2010) or Wilson (2019), the methods take as input series of test statistics (or p-values).

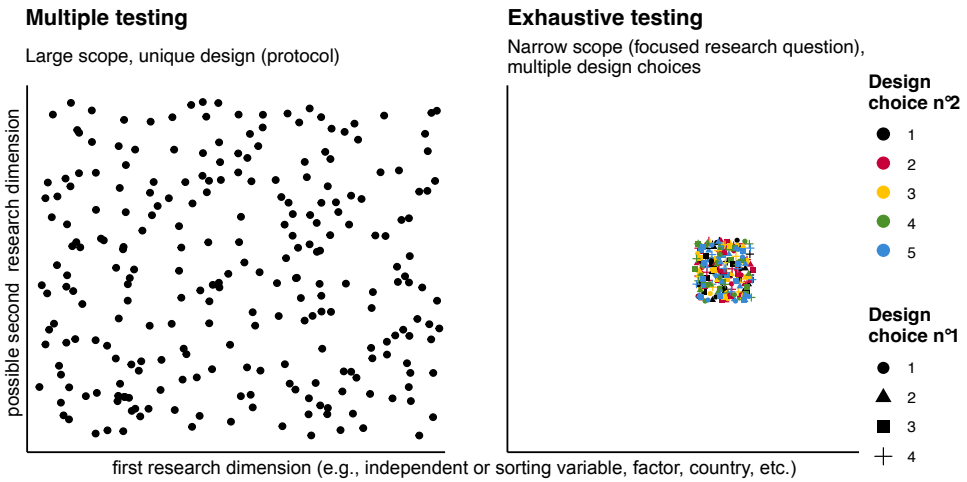

Figure 3: Illustration of **model rich and protocol poor** (many hypothesis, one path) versus **model poor and protocol rich** studies (one hypothesis, many paths).

that emanates from the data consistently estimates the true distribution of the effect. This is arguably a strong assumption which completely precludes sampling bias, which is why we propose an alternative procedure to generate more diverse samples.

Below, we formally recall the so-called *bootstrap reality check* (BRC) of White (2000), as it is exposed in Harvey et al. (2020). It will serve as benchmark, and the method can effortlessly be extended to the StepM procedure of Romano and Wolf (2005). We are given $T \times N$ observations $x_{t,n}$, where T stands for the sample size and N is the number of tests. These observations are bootstrapped B times to yield a $B \times T \times N$ tensor $x_{t,n}^{(b)}$. Here, b is the index of the bootstrapped sample. Bootstrapped statistics are defined as

$$t_n^{(b)} = \sqrt{T}\,\frac{\mu_n^{(b)} - \mu_n}{\sigma_n^{(b)}}, \tag{27}$$

where $\mu_n^{(b)}$ and $\sigma_n^{(b)}$ are the sample mean and standard deviation of each bootstrap series. μ_n is the sample mean of the original (non bootstrapped) data. We write $\tilde{t}_n^b$ for the statistics ordered such that $\tilde{t}_n^{(b)} \geq \tilde{t}_{n+1}^{(b)}$, so that, for each bootstrap sample b, $\tilde{t}_1^{(b)}$ 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 $\tilde{t}_1^{(b)}$.

Now, instead of resorting to bootstrapped samples, we propose to use forking paths to generate alternative versions of the same problem. The rationale is that one path is subject to sampling bias, whereas many paths provide a richer characterization of possible states of the world (different sub-periods, various weighting schemes, etc.). Consequently, we propose the following adjustment to Equation (27):

$$t_n^{(p)} = \sqrt{T_p}\,\frac{\mu_n^{(p)} - \mu_n}{\sigma_n^{(p)}}, \tag{28}$$

where bootstrapped samples b are replaced by paths p which have sample sizes T_p. In this case, μ_n can be the average over one representative path, or over any set of paths. Like for the bootstrapped statistics, the values are sorted to yield $\tilde{t}_1^{(p)}$.

The main difference between the two approaches is that, in the first case, $\mu_n^{(b)}$ is computed via the same data as μ_n, so that the numerator is expected to be smaller in magnitude compared to $\mu_n^{(p)} - \mu_n$ because $\mu_n^{(p)}$ will be based on possibly very different samples. One important question is whether the denominators will mitigate these differences so that $\sigma_n^{(p)}$ will shrink the dispersion of path outcomes.

The consequences can further be theoretically illustrated. Say we consider N Gaussian variables $t_n^{(\sigma)}$ which stand for anomalies' test statistics. For ease of exposition and analytical tractability, we assume that they are independent, with common zero mean and standard deviation σ. Then, we define the cdf of their maximum:

$$F_\sigma(x) = P\left[\max_{n \leq N} t_n^{(\sigma)} \leq x\right] = \Phi_\sigma(x)^N, \tag{29}$$

where $\Phi_\sigma(x)$ is the related Gaussian cdf. We are interested in the sensitivity of the quantile function (for the p-value, or significance threshold) with respect to σ, and for a fixed x, say $x = 95\%$:

$$\frac{\partial}{\partial \sigma}(F_\sigma^{-1})(x) = \frac{\partial}{\partial \sigma}\left(\sigma\sqrt{2}\,\text{erf}^{-1}(2x^{1/N} - 1)\right) = \sqrt{2}\,\text{erf}^{-1}(2x^{1/N} - 1) > 0 \quad \text{for } x > 2^{-N},$$

where erf^{-1} is the inverse error function. The above result intuitively proves that, as the dispersion of statistics increases, the threshold of their maximum will also increase. Given that we expect more dispersion from path-generated outputs, the resulting decision hurdles for significance should be larger. This will be investigated empirically and corroborated in Section 4.2.

The fact that path-generated decision thresholds will be more conservative implies that they will provide a stricter control over first type errors (false positives). However, this also means that they will be more prone to errors of the second kind (false negatives). The recent literature on multiple testing (Romano and Wolf (2005), Harvey et al. (2020), Harvey and Liu (2021a)) is very focused on this notion of **power**, i.e., the need to avoid as many false negatives as possible. This makes sense: controlling type 1 errors, while maximizing power amounts to seeking maximum accuracy and hence revealing a maximum number of *true* factors.

Nevertheless, there are often asymmetries between the consequences of both types of errors. In the money management industry, false positives incur losses (at least relatively to a benchmark), whereas false negatives are simply missed opportunities. Hence, arguably, investors may be more sensitive to false positives. Exhaustive multiple testing provides a stringent filter: anomalies that pass the threshold are immune to a large amount of variation in protocols and sub-sampling. Seeking this kind of robustness should be a prerequisite in portfolio backtesting. However, a drawback is that our approach is not well suited for purely inferential purposes.

2.6 Corroborating published results

Published results can be subject to bias towards positive outcomes and consequently researchers often seek to reproduce studies. Most of the time, perfect replicability is impossible because of data availability, interpretation leeway, and coding choices (see Pérignon et al. (2023) for more on the matter). Nevertheless, even if coefficients or t-statistics cannot exactly be reproduced, it is useful to evaluate if their magnitudes are reasonable - and forking paths are the appropriate tool for this task.

Indeed, once paths have been generated, it is interesting to compare the distribution of their outcomes to results of contributions that have quantified the same effects. Let us illustrate this with an example from our study in Section 3. In Figure 4 below, we plot the histogram of coefficients of predictive regressions when forecasting the market return with the aggregate book-to-market ratio, along 1,152 paths. From the distribution of the path-generated coefficients in the Figure, it is clear that the link between market returns and the valuation ratio is positive. Nevertheless, the

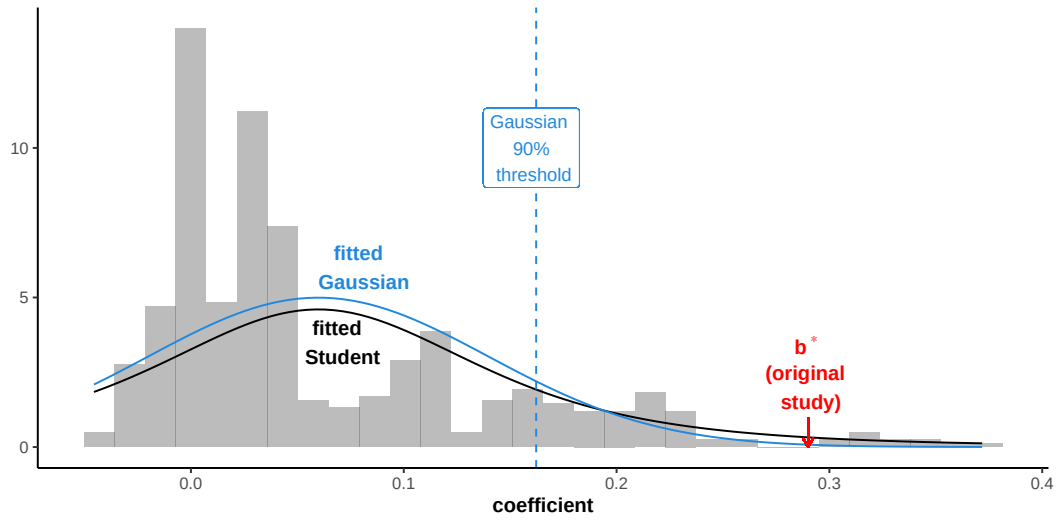

Figure 4: **Comparison with prior work**. We report the distribution of coefficients in a predictive regression exercise. The independent variable is the *aggregate book-to-market*. At the right of the histogram, we show the value ($b^* = 0.29$) from the study of Goyal et al. (2023).

value documented in Goyal et al. (2023) lies quite at the right of the distribution, meaning that it corresponds to some of the *best* outcome from our 1,152 paths.

In order to locate the position of outcomes from other studies within the distribution of path-generated results, we naturally propose to think in terms of quantiles. This way, the external results are assessed through the lens of their position within the paths' outcomes. If Φ_P is the empirical cdf of the path-generated outcome and if b^* is the outcome of reference from a prior study, then we define the probability level of the prior effect as $\Phi_P(b^*)$. It is the probability that the documented effect size b^* be larger - more extreme and favorable - than path values. Henceforth, we assume without loss of generality that the effect is positive and that we consider values to the right of the distribution to be favorable. For negative effects, we simply need to change the signs of results ex-ante.

For instance, if the probability $\Phi_P(b^*)$ is 0.6, then it means that the prior published result is close to the middle of the distribution obtained from the paths and thus there is little reason to suspect any cherrypicking because the reported value is so to speak *representative* of the plurality of outcomes. However, if $\Phi_P(b^*) = 0.99$, then the documented effect is more extraordinary. For example, in Figure 4, $\Phi_P(b^*) = 0.98$.

Nevertheless, we see that this approach faces a potential limit when the external value lies outside the range of path-generated results, which will occur in Section 3. To address this issue, we propose to resort to parametric continuous cdfs, which we will write $\hat{\Phi}_P$, as their parameters will be estimated from paths' outcomes. This is illustrated in Figure 4 with the blue curve which shows the fitted Gaussian density. Under this density, the result of Goyal et al. (2023) study is located at the 99.8% level. If we want to allow for more diversity in outcomes and increase tails, it is possible to switch to other distributions, like the Student law (in black in Figure 4 - with 3 degrees of freedom). In this case, the probability shrinks to 96.8%.

Based on these levels, our aim is to devise an indicator that evaluates the likelihood that a given outcome is *outstandingly* favorable. To do so, we focus on the empirical distribution of b, conditional on b being larger than some threshold effect θ. For instance, we pick θ to be at the 90% level of all paths-generated effects (or t-statistics) so that $\Phi(\theta) = 0.9$. This is shown with the vertical blue line in Figure 4. Plainly, θ is a large yet realistic value for b. We then define the **odds of**

favorable outcome (OFO) for b^* as the probability with respect to the right-truncated distribution as follows

$$\mathbb{OFO}_P(b^*, \theta) = \mathbb{P}[b < b^* | b > \theta] := \frac{\Phi_P(b^*) - \Phi_P(\theta)}{1 - \Phi_P(\theta)} \mathbb{1}_{\{b^* > \theta\}} \approx \frac{\hat{\Phi}_P(b^*) - q}{1 - q} \mathbb{1}_{\{b^* > \theta\}} \in [0, 1], \quad (30)$$

where, in the approximation, we have simply replaced Φ_P by its parametric proxy $\hat{\Phi}_P$ and $q = \hat{\Phi}_P(\theta)$ is a high benchmark level such as 0.9 for instance. Simply put, the OFO measures the likelihood of the reported effect to be above a value that is already considered large based on the path outcomes. Plainly, the OFO only makes sense for $\theta \leq b^*$. Moreover, it strongly relies on two key hypotheses. First, we assume that, as the number of paths increases, the average of the effects converges to the true value $\bar{b}$.[15] Second, the OFO will depend, via $\hat{\Phi}_P(b^*)$, on the standard deviation of the effects across the paths. In order not be too conservative against published results, we hope to report values that can only understate the true OFO obtained with the actual distribution of effects, Φ. This happens if the realized standard deviation of effects across paths is larger than σ_b, the true deviation.

A favorable feature of the OFO metric is that it can be *loosely* interpreted as a probability of original numbers being uncommonly favorable. Reversely, this can also be viewed as a measure of replicability: a small OFO signals a value that can be easily reproduced. For simplicity, we will henceforth report the EtC metric, which stands for "*Ease to Confirm*" or "*Ease to Corroborate*", and is equal to one minus the OFO: a value close to zero (*resp.* one) signals a result that is hard (*resp.* easy) to confirm.

Let us exemplify the indicator with the values from Figure 4. In this case, $\hat{\Phi}(b^*) = 0.98$ with a Gaussian proxy. We first fix $q = 0.9$ as extreme benchmark quantile and, consequently, EtC $= 1 - \frac{0.998 - 0.9}{1 - 0.9} = 0.02$ and the result from Goyal et al. (2023) stands out as clearly favorable and relatively hard to reproduce. If we impose an even more conservative threshold $q = 0.95$, then EtC $= 0.04$, a slightly higher score, but still underlining a figure that is hard to confirm.

Importantly, effect sizes can be both highly statistically significant *and* associated with EtC values that are very close to 100%, i.e., easy to reproduce. This corresponds to situations in which it is not arduous to confirm positive findings. We will substantiate this claim with further examples in Subsections 3.5, 4.3 and 5.4.

3 Application: equity premium prediction

3.1 Data

For the sake of reproducibility, the first illustration of the concepts of the paper rely on a public dataset as well as on a problem which is widely documented in the literature.[16] In financial economics, an old, still unresolved, question pertains to whether aggregate stock returns can be predicted by macro-economics indicators. The debate is likely impossible to settle, but recent results suggest a contingency on return horizon (Bandi et al. (2019)), even if long-term predictability is biased by construction for simple estimators (Boudoukh et al. (2008), Boudoukh et al. (2022)).

A critical view on the matter is the seminal article by Welch and Goyal (2008), in which the authors document the poor forecasting ability of traditional macro-economic predictors. A favorable feature of the study is that the data is public, and has even been updated in the follow-up

[15] This is a very technical point which is the subject of a separate paper, essentially based on Theorem 1 of Azriel and Schwartzman (2015).

[16] The code used to generate all results is available at https://www.gcoqueret.com/files/misc/forking_paths.html. The first version has been verified by the **cascad** certification service: https://www.cascad.tech/certification/116-forking-paths-in-empirical-studies/

paper Goyal et al. (2023). It is this material, updated until December 2021, that we use for our application.

For the sake of reproducibility, the first illustration of the concepts of the paper rely on a public dataset as well as on a problem which is widely documented in the literature.[17] In financial economics, an old, still unresolved, question pertains to whether aggregate stock returns can be predicted by macro-economics indicators. The debate is likely impossible to settle, but recent results suggest a contingency on return horizon (Bandi et al. (2019)), even if long-term predictability is biased by construction for simple estimators (Boudoukh et al. (2008), Boudoukh et al. (2022)).

A critical view on the matter is the seminal article by Welch and Goyal (2008), in which the authors document the poor forecasting ability of traditional macro-economic predictors. More recently, Dichtl et al. (2021) have confirmed the meager out-of-sample performance of most prediction methods. In addition, Engelberg et al. (2023) also report weak results when using aggregated cross-sectional indicators.

A favorable feature of the Welch and Goyal (2008) study is that the data is public, and has even been updated in the follow-up paper Goyal et al. (2023). It is this material, updated until December 2021, that we use for our application.

3.2 Forking paths

In order to generate enough metrics, we consider $J = 10$ stages (layers) which are depicted in Figure 5. We briefly comment on each below:

1. **data frequency** determines the horizon of returns, hence the left-hand side of the equation. In addition, this has a major effect on sample sizes, as annual samples are 12 times smaller, compared to monthly ones.
2. **handling missing points** boils to two options. The first is to remove rows of missing points, which means all regressors will start at the same point in time (1927). The second option (imputation of previous value) allows predictor-dependent sample sizes and some of them are available in 1871. Thus, this stage impacts sample depth but all the samples from the study have sizes above 40.
3. **winsorization** defines the cutoff threshold for the taming of outliers, from none (0%) to 3%. The data is often verified, thus all values are trustworthy, so this step could theoretically be omitted. But it participates to increase the number of outputs, hence we keep it for the sake of exhaustiveness.
4. **variable engineering** decides whether or not to use levels or differences in the regressions.
5. **independent variable** sets the predictor. Six options are possible and all are available across the three frequencies (monthly to annually).[18]
6. **horizon** fixes the number of periods that are used to compute the future return (dependent variable). We underline that three periods have different meanings depending on the original data frequency (chosen in step 1).
7. **starting point** determines if the sample commences at its first point, or at its middle point. This option leaves room for sub-sampling (on the two halves of each original sample).
8. **end point** is either the end of the sample, or its middle point. The latter option is not possible if it also corresponds to the starting point.

[17] The code used to generate all results is available at here. The first version of the code has been verified by the **cascad** certification service.

[18] **payout** is the difference between the log of dividends and the log of earnings, **b/m** is the the ratio of book value to market value for the Dow Jones Industrial Average, **svar** is the sum of squared daily returns on S&P 500, **dfr** is the difference between the return on long-term corporate bonds and returns on the long-term government bonds, **dfy** is is the difference between BAA- and AAA- rated corporate bond yields, and **ntis** is the ratio of twelve-month moving sums of net issues by NYSE listed stocks divided by the total market capitalization of NYSE stocks.

9. **estimation method** chooses between three specifications. First, the simple OLS with iid errors. Second, the improved HAC variance estimator of Newey and West (1987). The regression model for these two variations is simply[19]

$$y_{t+h} = a + bx_t + e_{t+h}, \tag{31}$$

where y_{t+h} is the equity premium (at horizon h), x_t the lagged predictor and e_{t+h} the residual. The third version is the augmented regression suggested in Hjalmarsson (2011):

$$y_{t+h} = a + bx_t + \gamma\nu_{t+1} + e_{t+h}, \tag{32}$$

where ν_t is the innovation process stemming from the predictor. More precisely, $\nu_t = x_t - \hat{\delta}x_{t-1}$, where $\hat{\delta}$ is the estimated coefficient from $x_t = \alpha + \delta x_{t-1} + \varepsilon_t$ (the constant term does not matter). The rationale for this is to treat potential endogeneity upfront, in addition to being very close to the method proposed by Amihud and Hurvich (2004) to solve the bias raised by Stambaugh (1999).[20] Finally, the fourth method combines the second and third (augmented regression with HAC variance estimation). The first two and last two have equal coefficients, but different t-statistics.

10. **post-treatment** seeks to correct potential error-in-variables bias. Once all paths have been generated, we have the choice follow or not Proposition 2 from Barras et al. (2022) to adjust the distribution of t-statistics.

There are $\prod_{j=1}^{10} r_j = 27,648$ possible paths from the data to the output. For simplicity, each is equi-probable so that we only need to consider each combination once.

3.3 Model averaging

In Figure 6, we show the averaged coefficients within their 95% confidence interval. We split the analysis along three axes: variable, level versus difference, and sampling frequency. The latter is important because it is determinant in the sample sizes which are used to compute the width of the intervals. For a given frequency and variable, they are homogeneous, though not exactly equal, and we use their weighted average $T_* = \sum_{j=1}^{J} w_j T_j$. In the figure, the impact of sampling frequency on the width of intervals is obvious. Intervals pertaining to monthly data coincide with the average estimators, whereas intervals linked to annual samples are fairly large. By definition, large sample shrink the interval ranges.

We observe that *ntis* yields only negative coefficients, and the intervals do not overlap with zero. The *b/m* variable has mostly positive estimates, with one exception of the quarterly data in the right panel. Surprisingly, the *dfy* variable also stands out with coefficients which are large in magnitude for quarterly and annual samples. For quarterly variables, the effect cancels out between levels (positive coefficients) and differences (negative ones). This partly explains why the variable was not previously identified as a potent driver of the equity premium.

In Figure 7, we plot the average coefficients computed according to Equation (21), along with ad-hoc confidence intervals. We only report results for annually sampled variables in order to reduce sample sizes. Indeed, their exponentiation in some term of Equation (24) are problematic when two models have significantly contrasting sample sizes. This issue is circumvented with annual samples.

The intervals in Figure 7 tend to confirm those obtained for the frequentist averages. Both *ntis* and *b/m* are associated with intervals that do not overlap over zero. In fact, for all predictors except one, the Bayesian averages are mostly indistinguishable from their frequentist counterparts.

[19] To keep scales comparable across horizons and frequencies, we scale the dependent variable by the square root of horizon time frequency. If the frequency is quarterly and the horizon 12 periods, we divide y by $\sqrt{36}$. Hence the baseline scale is the monthly return.

[20] The only difference is that instead of using $\hat{\delta}$, Amihud and Hurvich (2004) recommend to take $\hat{\delta} + (1 + 3\hat{\delta})/n + 3(1 + 3\hat{\delta})/n^2$, which is close to $\hat{\delta}$ whenever n is large enough, say $n > 40$.

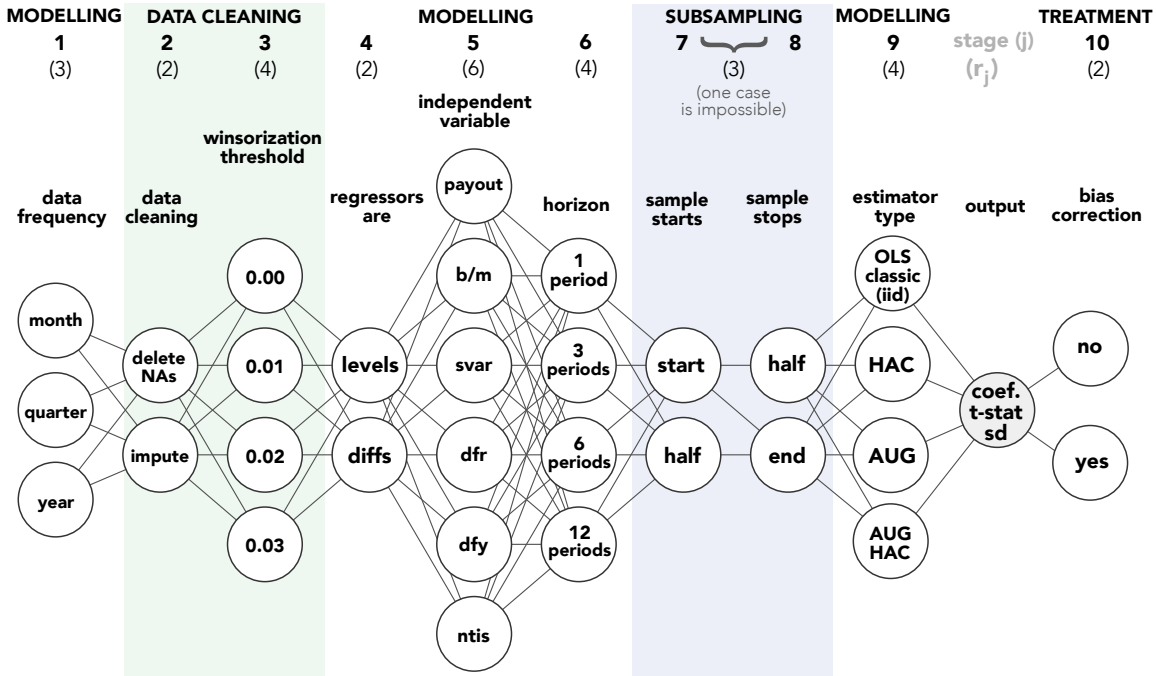

Figure 5: **Diagram of empirical protocol** (forking paths). The graph, akin to a neural network structure, depicts the ten-step algorithm used to produce the t-statistic in the study. Each path has the same probability of realization (uniform distribution). The number of mapping options (circles) is reported between brackets below the stage number. There are 27,648 paths in total.

3.4 Conditional averages

The above result show averages that are only conditioned by their sampling frequency. Below, we investigate the sensitivity of averages depending on other design choices, namely regression specification, subsampling, and estimated standard errors.

We start with the first two, jointly, to highlight the differences in the results. The two mappings, or layers we focus are the following. First, the estimation model, standard versus augmented, which corresponds to Equations (31) versus (32) in the definition of the layers. Second, we shed some light on the period-dependence of our results. Our protocol allows to look at subsamples and we are able to discriminate between the first half of samples (x-axis) versus their second half (y-axis).

In the top panel of Figure 8, we plot conditional means of standard models (Equation (31)) on the x-axis versus, on the y-axis the conditional means of coefficients obtained via the augmented models (Equation (32)). In addition, we draw the $y = x$ line to see where the points lie compared to the bisector. Points show the frequentist averages for the level predictors.

Plainly, switching from one to the other model specification has very limited impact, as almost all points are located very closely to the bisector. There is only one clear outlier, the quarterly *dfy* variable. When testing for the mean difference between the weighted values of the points (Equation (26)), the null of zero mean was rejected only once at the 5% confidence level (for *dfr* on annual samples) with a t-statistic of 2.0. All other statistics were below 1.9, even for the outlier point. In short, the model type does not affect results very much.

In the bottom panel of Figure 8, we proceed with the same analysis, but this time for period comparison. Given the strong time-variation in predictive coefficients (see Farmer et al. (2023)), we expect to see some difference in this case. And indeed, our results contrast with the upper panel

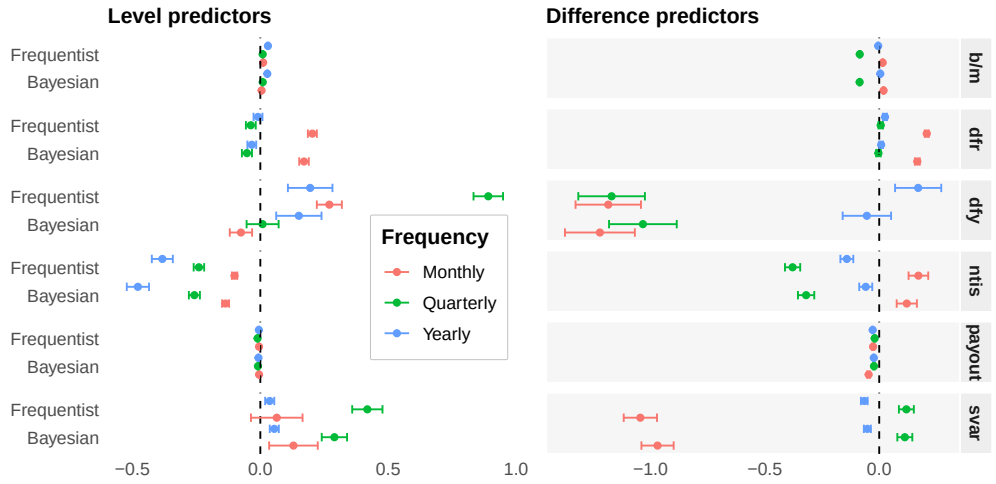

Figure 6: **Frequentist model averaging**. We display average coefficients within their 95% confidence interval. Coefficients stem from Equation (17) and predictors are scaled by their standard deviations prior to estimation to ease comparison of scales. Confidence intervals are defined by $[\hat{b}_* - 1.96\sigma_*^2/\sqrt{T_*}, \hat{b}_* + 1.96\sigma_*^2/\sqrt{T_*}]$, 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. Only the paths with no bias adjustment are considered.

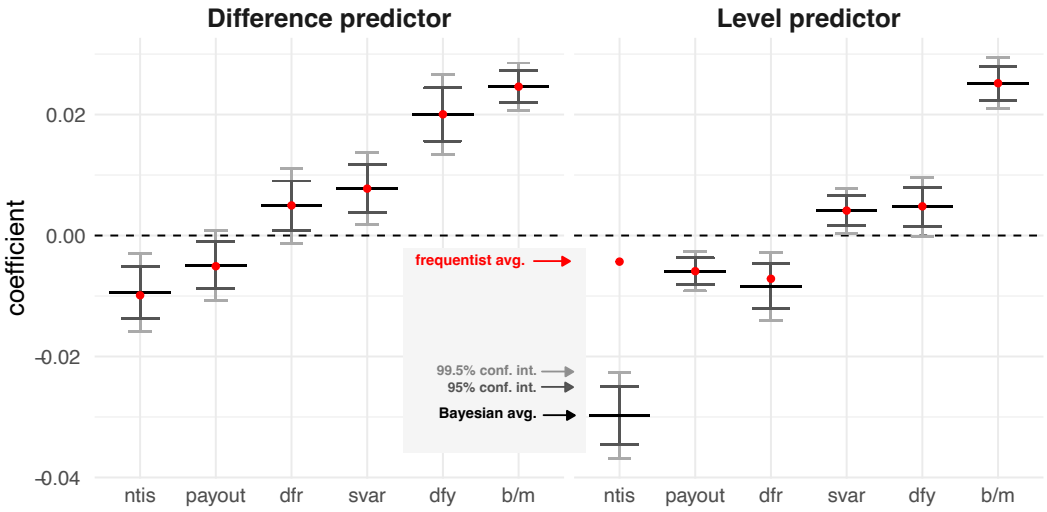

Figure 7: **Bayesian model averaging** (annual data). We display average coefficients within their 95% and 99.5% confidence intervals. Averages stem from Equation (21) and predictors are scaled by their standard deviations prior to estimation to ease comparison of scales. . The bounds of the confidence intervals are defined by $\mathbb{E}[b|D] \pm \alpha\sqrt{\mathbb{V}[b|D]/T_*}$, where $T_* = \sum_{j=1}^{J} w_j T_j$, with T_j being the sample size of model j and w_j the posterior model probabilities. α relates to the confidence level. The left panel displays results when predictors are differences, while the right one focuses on levels. To allow comparisons, all predictors are scaled to have unit variance before estimation. Only the paths with no bias adjustment are considered.

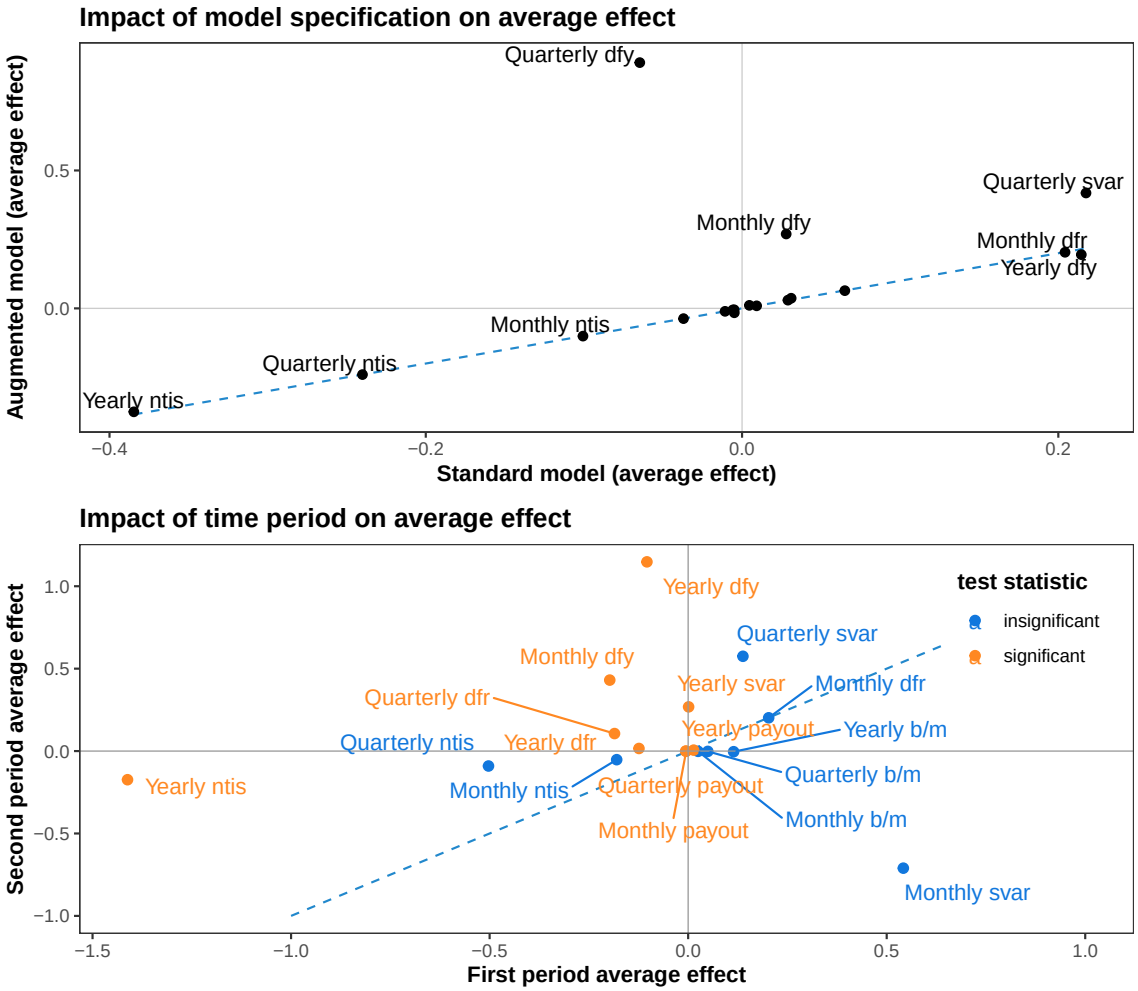

Figure 8: **Conditional differences**. We plot the conditional averages of one layer option (x-axis) version the other option (y-axis). In the top panel, the layer is the regression model (standard versus augmented) whereas in the bottom panel, the layer is the subsample (first versus second period). For both plots, we focus on level predictors only and the *frequentist* averages are obtained over 96 paths (top) and 128 paths (bottom). The dashed blue line marks the quadrant bisector. In the bottom figure, colors code for the significance of the simple t-test on the sequences (26) at the 5% confidence level.

because points no longer perfectly align with the bisector. The test statistics of the t-test defined in (26) and applied to the sequences of coefficients are split in two color categories depending on whether they are deemed significant at the 5% level. Surprisingly, there is no clear pattern for colors: the points closest to the bisector are not necessarily blue. Upon verification, this comes from large standard deviations which shrink the average effect in the denominator of the statistic in some cases.

Lastly, out of curiosity, we evaluate the impact of the standard deviation specification on the weighted average of t-statistics. In the above results, t-statistics are computed at the very end by taking the ratio between average effects and their standard deviations (19). This does not ac-

count for the model-specific estimates of standard deviations. In order to embed the latter in an aggregate measure, it is possible to average the t-statistics instead of raw effects.

In Figure 9, we produce the absolute value of weighted averages in test statistics. The y-axis features the HAC-corrected statistics. Therefore, because HAC standard errors are usually more conservative (i.e., larger), we would expect that the corresponding statistics be *smaller* than those from the i.i.d. estimator. And indeed, this is what we see in the plot, as almost all points lie below the dashed bisector. It is noteworthy to underline that the three points farthest to the right pass decision thresholds with the iid estimator and not with the HAC estimator. For instance, the *Yearly b/m* (*resp.*, *Monthly b/m*) predictor is significant at the 5% (*resp.* 1%) level with the iid estimator, but not with the HAC estimator.

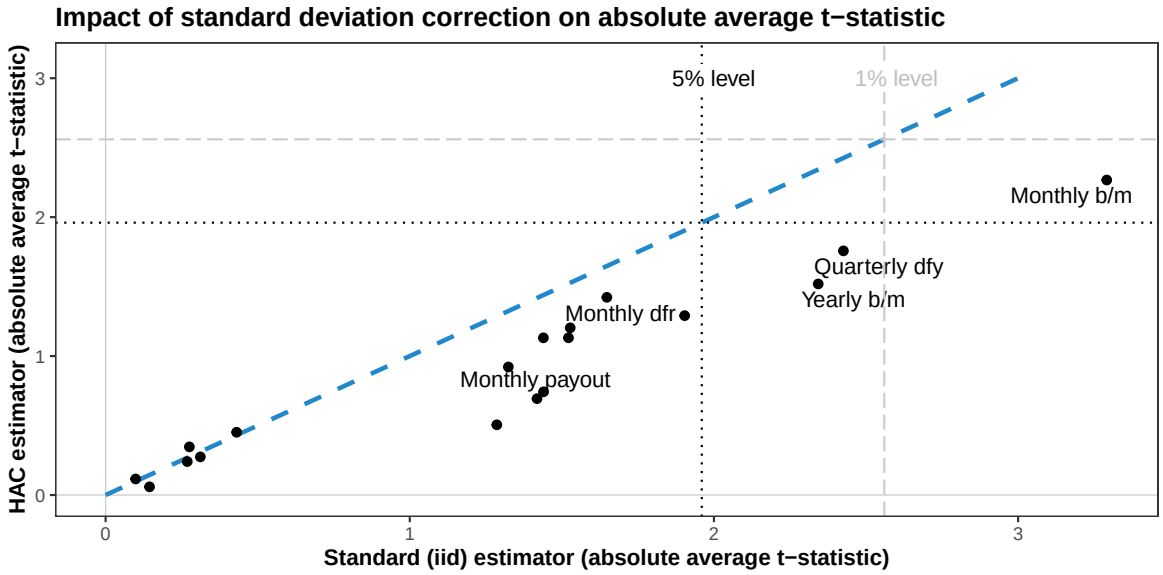

Figure 9: **Standard deviations and *t*-statistics**. We plot the absolute weighted averages of t-statistics (the absolute value is applied after the averaging). The x-axis relates to the average statistics computed with uncorrected standard deviations, whereas the y-axis pertain to statistics that adjust for auto-correlation and heteroskedasticity. Vertical and horizontal dashed lines mark the decision thresholds at the 5% and 1% levels.

3.5 Comparison with other studies

In Goyal et al. (2023), the authors report the coefficient estimates from their regressions, both on their full sample and on the two halves of their initial sample. This is in contrast with Dichtl et al. (2021) and Engelberg et al. (2023) who do not provide the raw estimates, but only out-of-sample fit. Henceforth, we focus on the monthly values provided by Goyal et al. (2023), which also correspond to level predictors, and not difference variables (increments).

In Figure 10, we depict the distribution of the path-generated values, alongside those from Goyal et al. (2023). There are three of them: the full sample in black and the two halves in red. For four predictors (all but *b/m* and *payout*), the premia over the full sample are all well within the values obtained from the paths and they can hence be reproduced effortlessly. For the *b/m* predictor, it seems the valued reported by Goyal et al. (2023) is slightly optimistic. However, for *payout*, there appears to be a scale problem - though we stuck to the variable definition in the original paper.

For the premia related to the sample halves, the ease to corroborate depends on variables, but paths are mostly compatible with at most one of the halves (e.g., having a EtC score above 50%). Our results also underline, somewhat as expected, that large sample sizes are clearly associated with the highest generalization ability. Effect sizes computed on deep chronological samples are less prone to historical idiosyncrasies.

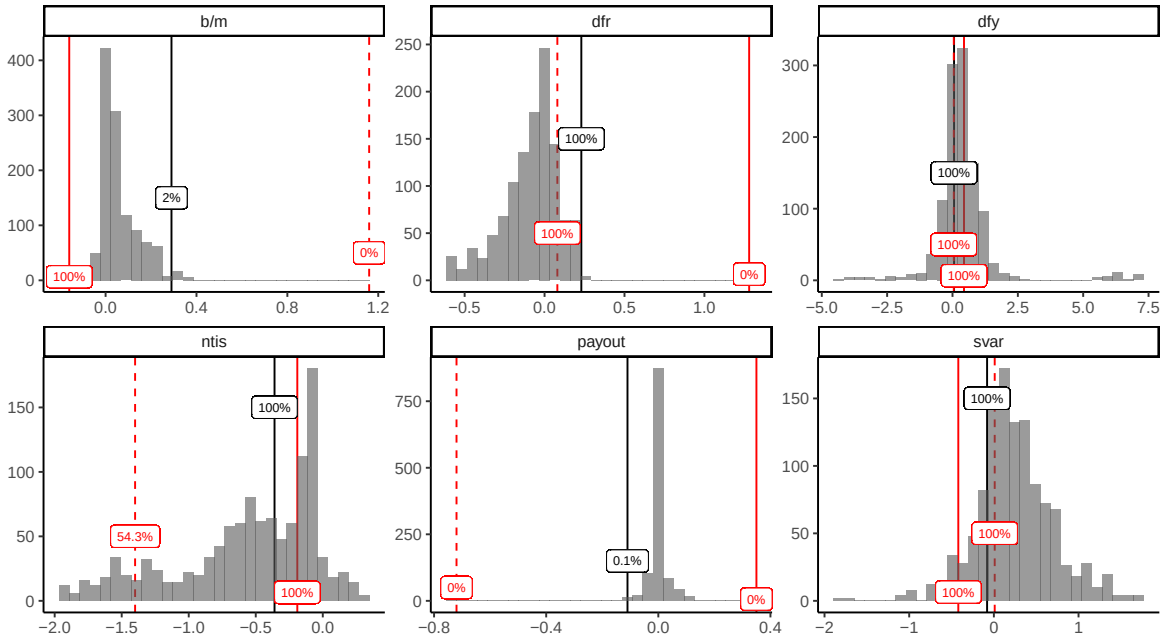

Figure 10: **Comparison with Goyal et al. (2023)**. We display the distribution of effects across paths with grey histograms. The black line shows the value obtained by Goyal et al. (2023) over the full sample. The red lines represent the values for the two halves of the sample (dotted for the first half and full for the second half). The rounded rectangles provide the the *ease to confirm* indicator (EtC) with respect to the Gaussian distribution (with mean and standard deviation computed from the histograms). It is equal to one minus (30) with $q = 0.9$.

3.6 Expansion rates of hacking intervals

The large number of paths generated in the study allows to compute the quantities defined in Equations (14), (15) and (16), when the outcome is the t-statistic. In Figure 11, we show the box-plots of interval ranges proposed in Equation (14). In the case when the x-axis $J - k = 1$, to the left of the plot, there is only one free mapping, which generates small dispersion in t-statistics. This is why both the median and average of the ranges is small (below one). As one shifts to the right of the plot, more and more leeway is given to the researcher and the breadth of output increases. If we assume a constant rate of increase, we obtain (via log least-square optimization) that each mapping expands average intervals by 42% (which we see with the black power line).

However, it is clear that the effect is not uniform: the line is below in the middle, but above towards the end. This signals the intuitive pattern that as $J - k$ increases, the speed of expansion slows down because intervals are already very large. This is confirmed in Table 2 below, which provides speed at which the average intervals increase. As expected, the increase rate, ρ_{J-k} defined in Equation (16), decreases with $J - k$. Nonetheless, the final value (after 9 successive increments) remains close to 30%.

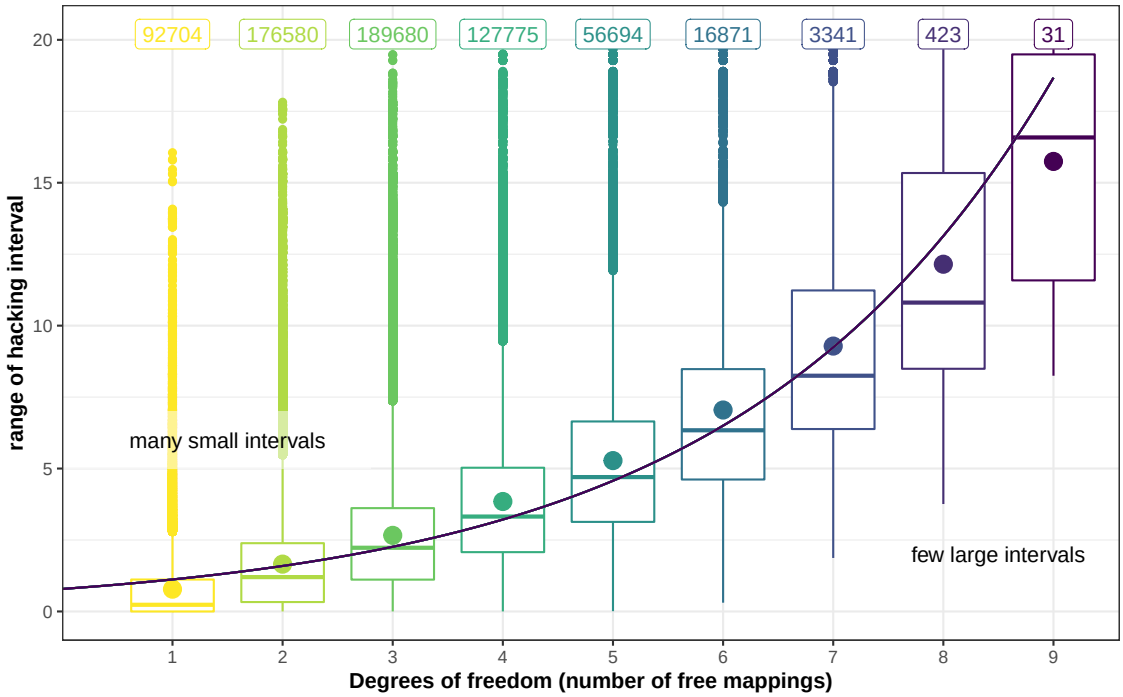

Figure 11: **Expanding intervals**. We show the boxplots of hacking interval ranges defined in Equation (14)). The small vertical points stand for outliers. The bigger circles represent the mean of ranges (ARI), introduced in Equation (15). The curve is parametrized by: $ARI(n) = 0.78 \times 1.42^n$, where the constant were estimated by least square minimization to feat the mean points in a log-regression ($\log(ARI) = \log(a) + \log(b) \times n$). The numbers at the top provide the number of intervals available for each number of free mappings.

$J-k$	1	2	3	4	5	6	7	8	9
ARI	0.779	1.656	2.662	3.847	5.278	7.048	9.285	12.149	15.745
rate (ρ_{J-k})		1.124	0.608	0.445	0.372	0.335	0.317	0.308	0.296

Table 2: **Interval expansion**. We report the average range of intervals (ARI, from Equation (15), along with its growth with the number of free mappings ($J-k$).

The extreme case ($J-k=9$) occurs when only one mapping is fixed (each option being considered separately) and this is highlighted in Figure 24 in Appendix D. There are 31 alternatives and the narrowest interval is the fourth one (*dfy*), with a width close to 8, which is the lowest point in the rightmost candlestick in Figure 11. Several choices yield intervals with range close to 20 (from -9 to +11), which is the upper limit of the candlestick.

All in all, our results confirm the self-evident idea that, as researchers consider more ways to run their protocol, they should expect a wider range of outcomes. Ideally, this would result in a better characterization of the effect they study.

4 Application: anomalies from portfolio sorts

The literature in financial economics has seen a surge in *factors* (see Harvey et al. (2016)). The latter are also called *anomalies* because they contradict the cornerstone result that is the capital asset pricing model (CAPM). The multiplication of publications in the field has even led researchers to devise new tests and approaches to detect or analyze when a factor is truly a factor (e.g., Feng et al. (2020), Chinco et al. (2021) and Harvey and Liu (2021a)).

The simplest way to proceed, since the seminal work of Fama and French (1992), is to periodically sort stocks according to some characteristic and test if extreme quantile portfolios have significantly different means. This again gives rise to implementation leeway, such as holding periods, quantile thresholds, and portfolio weighting for instance. This section is therefore dedicated to the impact of these modelling choices on the significance of asset pricing anomalies. Its main goal is to illustrate the concept of exhaustive multiple testing presented in Section 2.5. Our conclusions on the sensitivity of design choices corroborate some findings in the recent similar studies of Bessembinder et al. (2022), Soebhag et al. (2023) and Walter et al. (2023).

4.1 Data and paths

We rely on the dataset used in Gu et al. (2020) updated by the authors until the end of 2021. In the original study, 94 characteristics are used to predict returns, but some of them have limited support (e.g., they are binary), and are not suitable for sorting purposes. We remove these variables and are left with 82 characteristics.[21] After this filter, some changes in data availability occur prior to 1983, hence we focus on the period from January 1983 to December 2021.[22]

In Figure 12, we depict the steps (modelling choices) that we consider to qualify asset pricing anomalies. In total, there are 576 paths for each sorting variable, which makes 82×576=47,232 paths in total.

More precisely, the seven modules are:

1. the **sorting variable**;
2. the **data cleaning** choice (imputation of prior value or removal);
3. the **holding period** posterior to sorting (1, 2 or 3 months - this corresponds to reasonable choices for rebalancing frequency);
4. the **starting point** of the study (minimum, first third or second third of the full sample's dates)
5. the **ending point** of the study (first third, second third, or maximum of the full sample's dates). This means that the smallest samples have a length that is one third of the total sample (∼39 years), hence encompassing several macro-economic environments.
6. the **quantile threshold** (q) used to compute the long-short portfolios (long the upper $1-q$ stocks and short the lower q stocks. The factors' sensitivity to breakpoints is thoroughly investigated in Hollstein et al. (2022).
7. the portfolio **weighting scheme**: equally-weighted (EW), inverse volatilty-weighted (IVW) and value-weighted (VW). With IVW, weights are proportional to the inverse of the *retvol* characteristic and with VW, they are proportional to the *mvel1* indicator. Finally, we also consider an alternative scheme, CW (characteristics-weighting), for which the weight in the

[21] The full list of abbreviated names is: absacc, acc, aeavol, agr, baspread, beta, betasq, bm, bm_ia, cash, cashdebt, cashpr, cfp, cfp_ia, chatoia, chcsho, chempia, chinv, chmom, chpmia, chtx, cinvest, currat, depr, dolvol, dy, ear, egr, ep, gma, grcapx, grltnoa, herf, hire, idiovol, ill, indmom, invest, lev, lgr, maxret, mom12m, mom1m, mom36m, mom6m, mvel1, mve_ia, operprof, orgcap, pchcapx_ia, pchcurrat, pchdepr, pchgm_pchsale, pchquick, pchsale_pchinvt, pchsale_pchrect, pchsale_pchxsga, pchsaleinv, pctacc, pricedelay, quick, rd_mve, retvol, roaq, roavol, roeq, roic, rsup, salecash, saleinv, salerec, secured, sgr, sp, std_dolvol, std_turn, stdacc, stdcf, tang, tb, turn, zerotrade.

[22] The code and data used for this part can be accessed online.

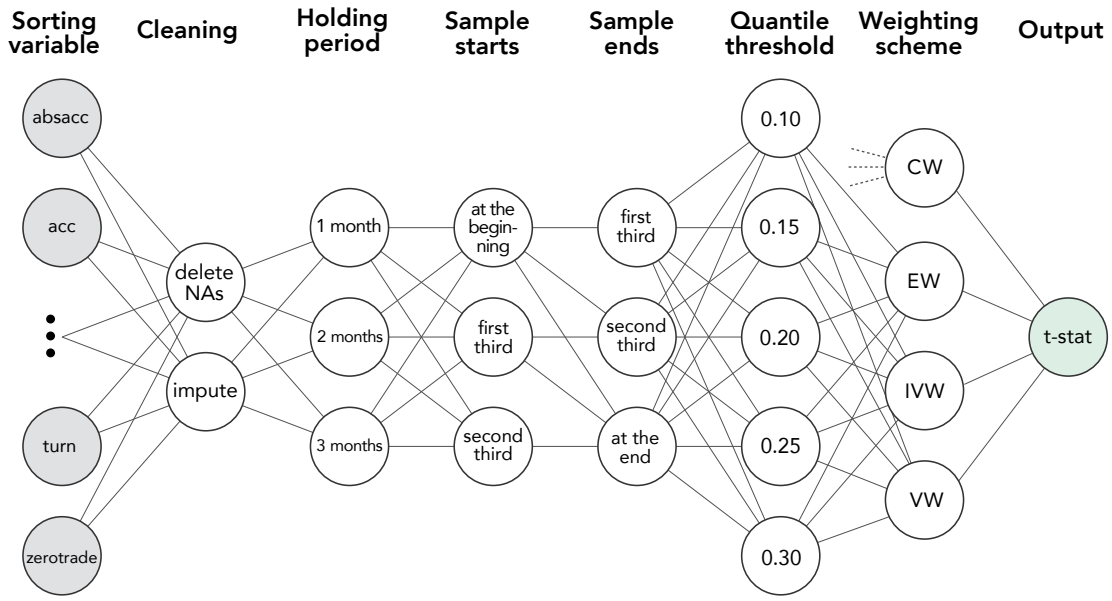

Figure 12: **Forking paths for anomalies**. The graph, akin to a neural network structure, depicts the seven-step process used to produce the t-statistic (last layer) in the study. Each path has the same probability of realization (uniform distribution).

long-short portfolio is the scaled value of the characteristic such that each leg has weights that sum to one. In this case, weighting is smoother in the cross-section, as there is no threshold.

In addition, we define a default, or baseline, path:

► **default-representative path** (parametrization of the layer options): imputation of missing data (and not deletion), a one month holding period, 0.2 sorting threshold (quintiles), equally-weighted portfolios, and the average returns and statistics are computed on the full sample.

Each path produces a series of portfolio returns. The output corresponds to a simple t-test on the mean of the average return: $\sqrt{T}\mu/\sigma$, where μ is the average and σ the standard deviation of returns. Up to the size scaling and assuming zero risk-free rate, the output is the Sharpe ratio of the portfolio.

These outputs are summarized in Figure 13. Notably, the width between the extreme points of the boxplots correspond to the range of hacking intervals for each factor. An anomaly may be considered to be strong if these intervals are not centered around zero. A remarkable pattern is the contrast between the short-term reversal (negative returns stemming from one month momentum) and the one year trend-following strategy which is associated to positive returns. The variety of results suggests that some anomalies are definitely more robust than others.

Half a dozen factors are able to sustain positive average returns over the full wide scope of implementations: *roic*, *roeq*, *retvol*, *mom6m*, *maxret* and *cashdebt*. Because of the multiple environments in which they have been tested,[23] they emerge as robust strategies within the zoo of factors. Our results are yet another confirmation of the momentum factor, which is widely documented as a persistent one (see Smith and Timmermann (2022) for a recent appraisal).

[23] Formally, we cannot consider that each path represents an environment of its own. However, two paths (for a fixed characteristic) with no overlap in layer options may be considered as two complementary facets of the same factor.

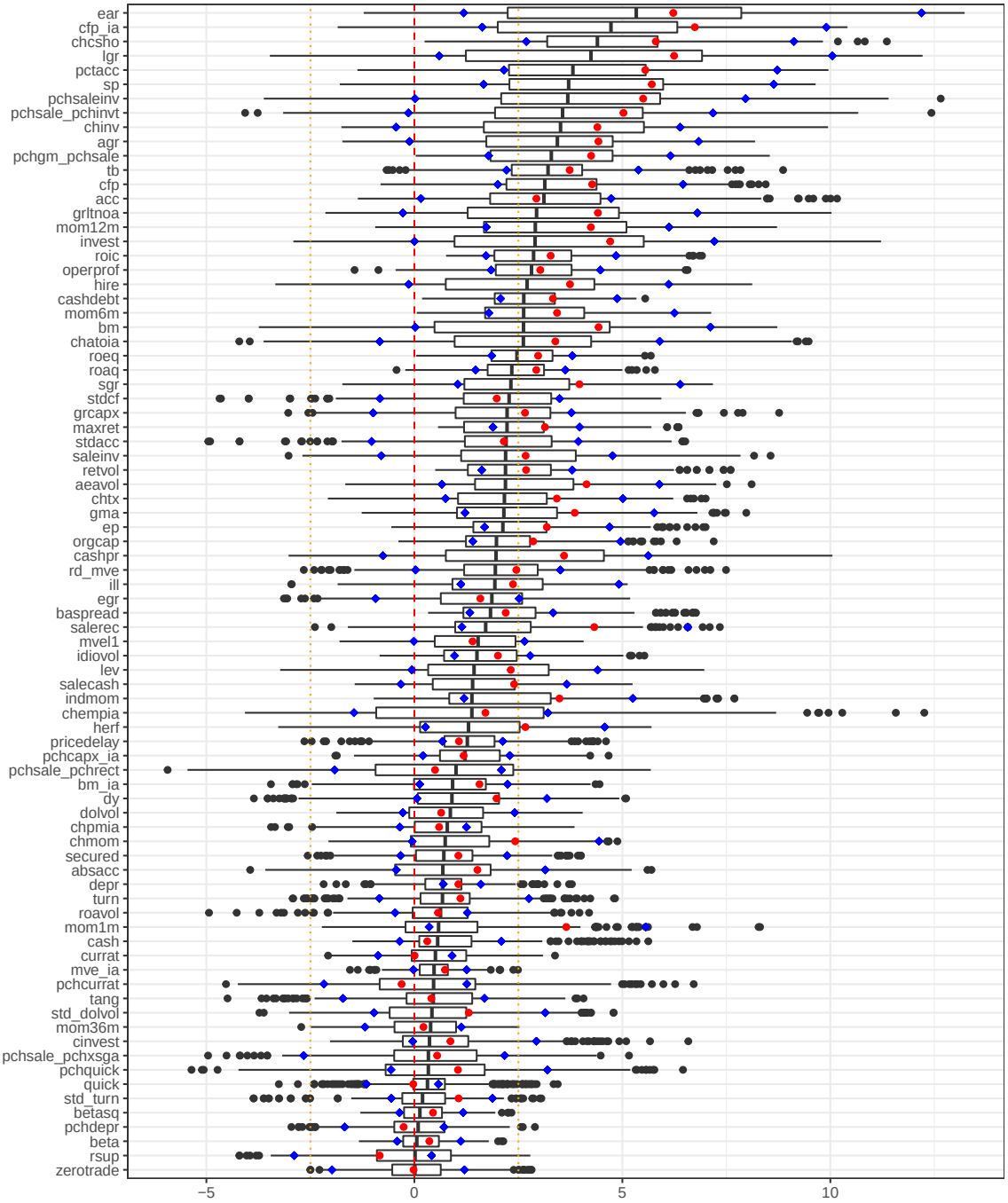

Figure 13: **Distribution of output**. We produce the boxplots (over 576 paths) for the t-statistics for the simple test of significance of average returns of each of the 82 anomalies (shown on the y-axis). Variables are ordered by median t-statistic and for each variable the sign (long versus short) is adjusted so that this median be positive. The red point marks the statistic for the baseline path. The blue diamonds mark the range of statistics when paths are restricted to robustness checks (14 paths).

In addition, in Figure 13, we also show with blue diamonds the extreme points of what we refer to as *robustness checks*. These are all the paths which have exactly one deviation from the default path. Given the path structure depicted in Figure 12, this makes 14 paths. The average range, across anomalies, for intervals of t-statistics is 4.1 for robustness checks, while it is 9.1 if we span all the paths. This forcefully shows that simple robustness checks only provide a narrow picture of the diversity in outcomes, whereas paths are much more exhaustive.

4.2 Multiple testing

Asset pricing anomalies are a fertile ground for multiple testing. The aim here is to test if factors are genuine, or simple flukes (see, e.g., Chen and Zimmermann (2022a), Harvey and Liu (2021a), Jensen et al. (2021) and Chen et al. (2023)). We thus want to illustrate the exhaustive approach we advocate in Section 2.5. In this setting, the observations $x_{t,n}$ are the time-t returns of factor n.

We wish to compare the bootstrap reality check (BRC) approach at computing t-statistics (Equation (27)) to the exhaustive one that relies on paths (Equation (28)). To ensure comparability of the two methods, we rely on a number of bootstrap samples that is equal to the number of paths in the study (576). Moreover, we resort to block bootstrapping with blocks of size 12, i.e., coinciding to annual series.

In Figure 14, we plot the distribution of maximum statistics obtained from bootstrapping returns versus forking paths. Plainly, the distribution of bootstrapped statistics lies to the left of those stemming from forking paths. The main reason for this is that alternative paths generate average returns that vary considerably, compared to the baseline case. One particularly important layer is the subsampling one because anomalies can evaluated over different time-frames. While this may seem odd, it makes sense from an investment standpoint: a reliable long-short strategy should perform similarly over various periods, as long as these periods are long enough to encompass a variety of market conditions (e.g., bull and bear markets).

The main takeaway from this exercise is that by examining a large spectrum of outcomes instead of bootstrapped returns, extreme cases become more likely, which raises the bar for significance. With the default path, the maximum statistic for anomalies is 6.74 (for the *cpf_ia* characteristic, see rightmost red point in Figure 13), which is quite high. The hurdle at the 95% level from bootstrapping is 4.5, so that the best original anomaly passes the test handily. For the sake of completeness, we have tested the case with 10 times more bootstrap samples (5,760) and the threshold remains the same, at 4.5.

However, if we consider thresholds generated by paths, we obtain 8.2 or 12.7, depending on the configuration. They are the vertical dotted lines in the figure. In both cases, the baseline path of *cpf_ia* is not longer significant. In Chen et al. (2023), the anomalies that cannot be matched by data mining techniques all have t-statistics above 5, and a handful of them are even above 12. When a baseline statistic is very high, it is likely that paths in its vicinity will also have large values, often above 2 or 2.5. In Figure 14, we see that the variables for which the red point is far to the right (at the top) all have inter-quartile ranges do not include zero. For the corresponding accounting or risk characteristics, this means that 75% of the paths at least lead to profitable strategies. Hence, raising the threshold helps immunize anomalies against implementation sensitivity and false positives.

Naturally, the symmetric cost of this is a substantial increase of the odds of false negatives. A higher decision threshold means that most anomalies will not be able to reject the null of zero return. Therefore, it is inevitable that genuine factors be missed. From an inferential standpoint (e.g., for the researcher in asset pricing), this is a severe limitation. In contrast, for a portfolio manager false positives correspond to where the money goes, which is arguably what matters most.

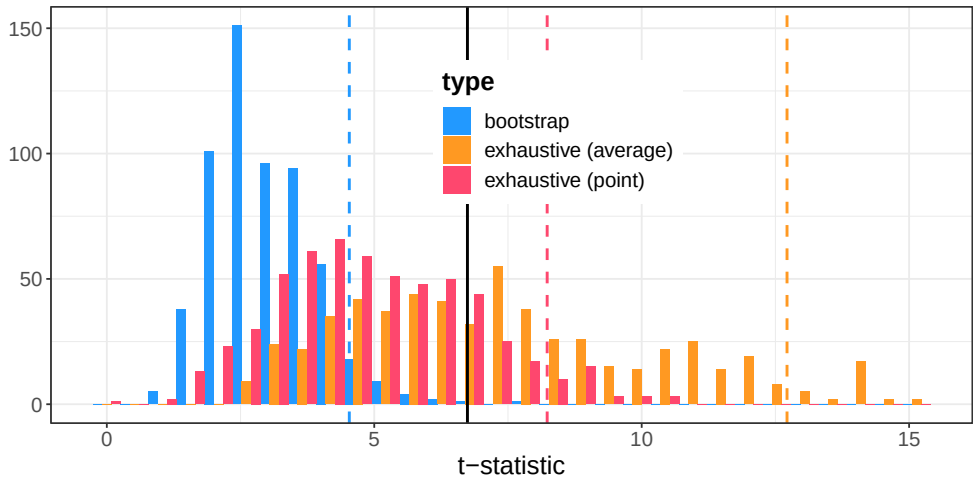

Figure 14: **Distribution of bootstrapped and paths-related maximum statistics**. We produce the histogram of the maximum statistics stemming from bootstrapping ($\hat{t}_1^{(b)}$, in blue) and forking paths ($\hat{t}_1^{(p)}$, in orange and red). The sequences are derived from equations (27) and (28), 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.

4.3 Comparison with prior work

The recent paper Chen and Zimmermann (2022a) provides an interesting basis to work with, as it lists and reproduces a large array of prominent asset pricing anomalies. Given the authors' work,[24] it is possible to compare the average returns (and the corresponding t-statistics) of sorted portfolios tested in prior studies. When identifying the academic articles from Chen and Zimmermann (2022a) with the sorting variables we worked with, we are able to determine an overlap of 40 predictors for which both returns and t-statistics are documented. For a few of them, several results are reported and we only keep the one that corresponds to the smallest average return, i.e., the one that will maximise the EtC indicator (ease to corroborate).

For each predictor, we can compute, given all average returns from the paths, the mean and standard deviation of these returns. It is then possible, as discussed in Subsection 2.6, to evaluate the position of the results from the former studies, if we assume a Gaussian distribution of outcomes from the paths. This is shown in Figure 15. Therein, the average returns of paths are shown as grey points and the confidence intervals (20) are so narrow that they are in fact represented with black vertical segments. The figures from original values reported by Chen and Zimmermann (2022a) are represented with red triangles.

Then, we compute the sample mean and variance of the grey points for each anomaly and report the EtC defined as one minus the indicator (30), where b^* is the red triangle value and Φ_P is the Gaussian cdf with mean and variance equal to their sample estimations. The EtC is reported at the right of the plot. Of the 25 anomalies we consider, 10 have an EtC above 50%. For these studies, the reported effects may be large, but they are not suspiciously so.

In all of these 10 studies, the reported values are statistically significant, meaning that, based

[24] They provide a CSV file that compiles the information from the literature.

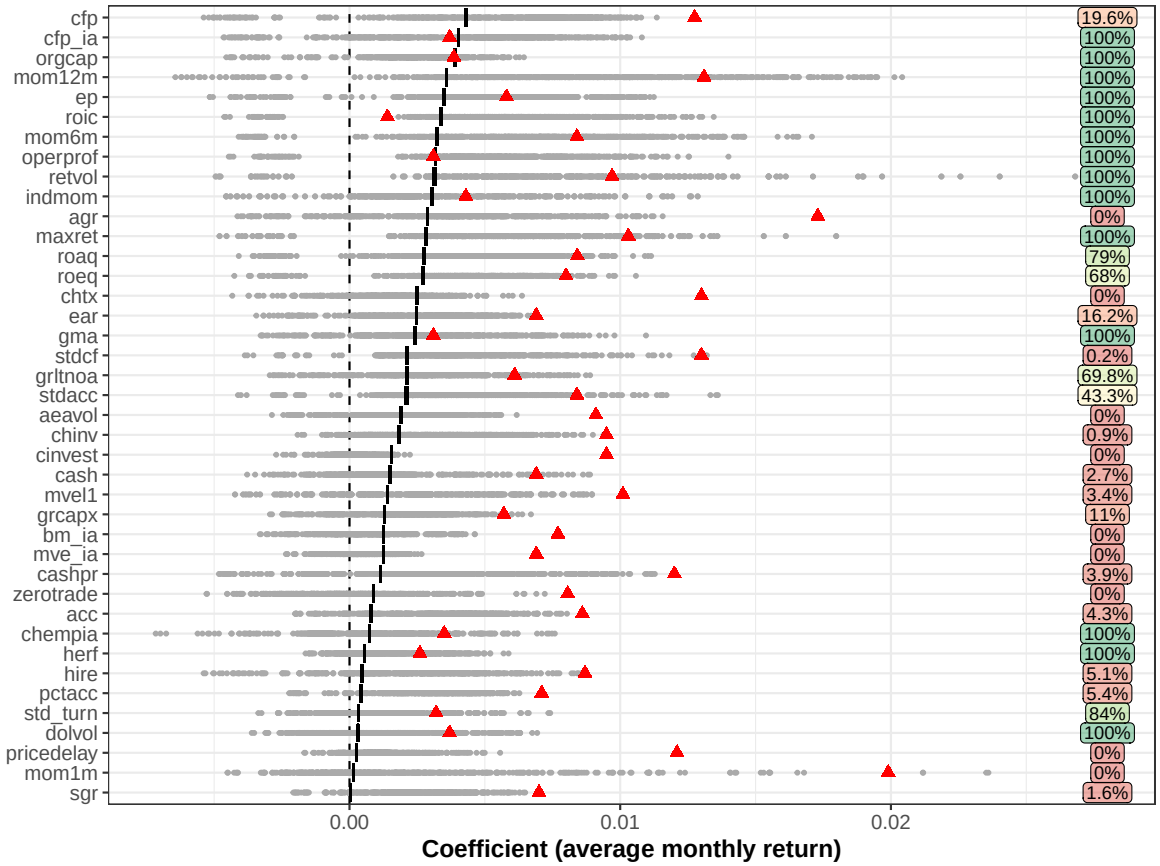

Figure 15: **Comparison with prior studies**. We display the average returns obtained by the 576 paths with small grey circles for each predictor (y-axis). The very narrow confidence intervals defined in Equation (20) are shown in black. Anomalies are ranked according to the average with weights (18). The average returns reported in the original studies are depicted with a **red** triangle. The values come from Chen and Zimmermann (2022a). At the very right of the plot in rounded rectangles, we provide the ease to confirm (EtC) of the triangle with respect to the Gaussian distribution fitted on the grey circles (mean and standard deviation) - equal to one minus (30) with $q = 0.9$. The colors code the facility to reproduce the original outcome.

on the paths, these results are strong and *likely* so. For instance, if we take the case of the *maxret* factor, in the original paper by Bali et al. (2011), the 1.03% monthly return is associated with an absolute t-statistic of 2.83, which makes it significant even at the 1% level. Therefore, this effect size is both significant and plausible.

On the other side of the spectrum, we report that 11 anomalies have an EtC below 10%, meaning that they correspond to values that were hard or even impossible to reach with the paths that we have spanned. This confirms that all anomalies are not equal. Some of them have returns that can sustain small alterations in the construction process, others do not.

4.4 Conditional averages: stability through time

Another way to test the potency of factors is to evaluate the shift of their performance through time. One convenient layer of the above protocol allows to split samples into three periods of equal sizes. This constitutes fertile ground to further illustrate the concept of conditional averaging

introduced in Section 2.4.2. The rationale for this is to evaluate if the average returns of sorted portfolios varies substantially across large non-overlapping periods. This is a major concern for money managers because a shift in one anomaly's profitability is likely to alter performance for agents who have invested in this particular anomaly.

In Figure 16, we plot average returns of anomalies across two different periods, one for each axis. In the upper panel, we compare the first period (1983-1996) to the second one (1996-2008) and in the lower panel, the second period is linked to the third one (2008-2021). Clearly, the relationship between the periods is stronger in the upper panel: this means that an investor in 1996 would have not been too disappointed by the performance of anomalies until 2008. However, there seems to be some decoupling posterior to 2008, as shown by the evidence in the lower panel: the correlation between the periods shrinks from 83% to 58%. Importantly, there are significantly more anomalies in the upper left and lower right quadrants, meaning that performance has reversed between the two periods. Nonetheless, there are some similarities and a few extreme anomalies remain in the same zones of the plots: *mom12m* (lower left) and *retvol* and *baspread* (upper right) emerge as stable factors. Only time will tell if this holds in the decades to come.

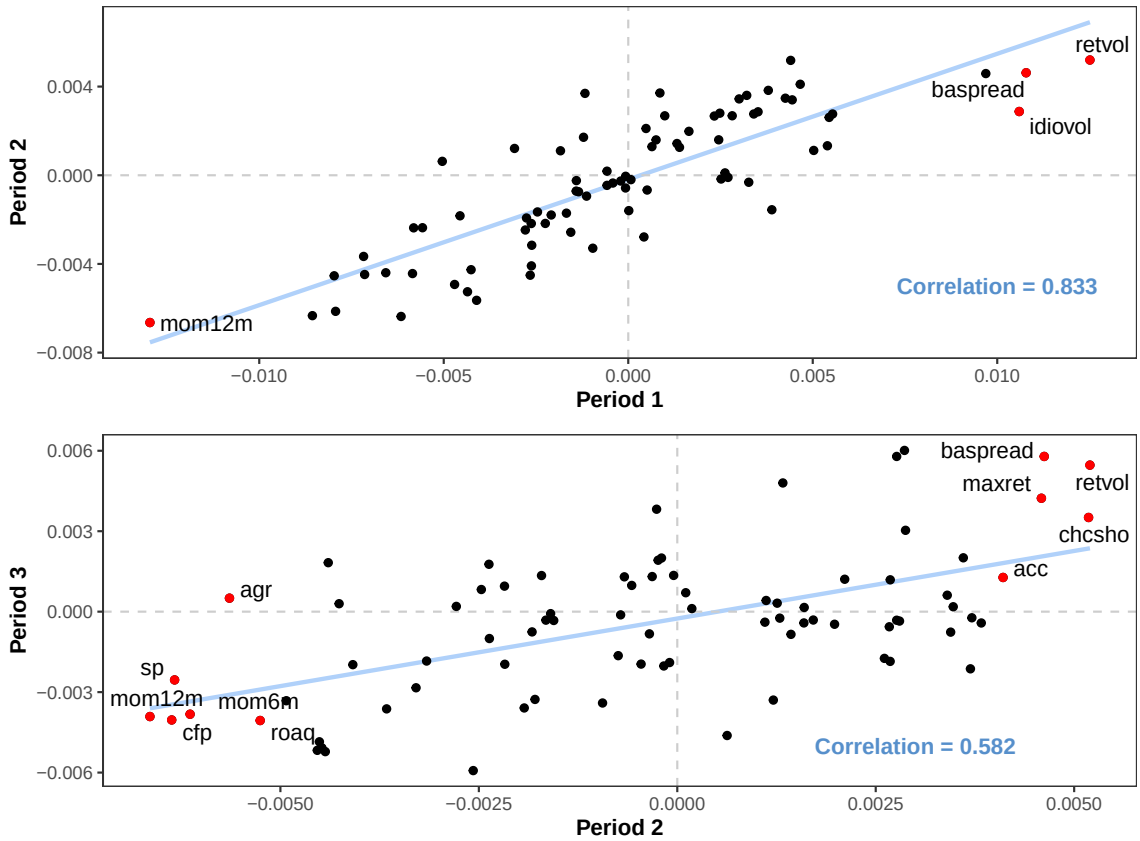

Figure 16: **Consistency through time**. We depict the simple average returns of sorts from the first to the second period (upper plot) and from the second to the third (lower plot). Each point pertains to one sorting variable and corresponds to the average over 96 paths. The light blue line shows the linear relationship fitted on all points. Extreme points are shown in red and the corresponding anomalies are specified.

4.5 p-hacking in the cross-section of characteristics

Elliott et al. (2022) provide criteria on the distribution of p-values to determine if a sequence of outcomes was generated by p-hacking. Surprisingly, as discussed in Appendix C, it may occur that even if p-values were not hacked, they fail to pass a test of no p-hacking. Therefore, after generating paths, it is interesting to determine how much falsification would be required to pass the most elementary p-hacking detection test. Because we only have 576 paths for each sorting variable, histograms may be somewhat noisy, which makes precise tests troublesome.

Instead, we resort to an ad-hoc classification of anomalies which is based on a weighted average of the decreasing rate of the p-curve. Namely, for each sorting characteristic, we compute

$$\kappa = \sum_{i=1}^{I/2} \frac{n_{i+1}}{n_i} \times \frac{1}{i}. \tag{33}$$

Without much loss of generality, we are assuming that the number of intervals in the histogram over the unit interval is even, hence all relevant frequency counts n_i are spanned for i between 1 and $I/2$ (the critical zone is $[0, 1/2]$). The discounting factor i^{-1} gives more weight to the first bars of the histograms because after a few rounds of decrease, the values of bar counts is very noisy and ratios are less trustworthy. Our classification is as follows:

- p-hacking is **unnecessary** if $\kappa < 0.25$.
- p-hacking is **possible** if $\kappa \in (0.25, 0.4]$.
- p-hacking is **problematic**, otherwise.

Of course, these choices are arbitrary and determined ex-post, but, as is shown in Figure 17, they yield coherent groups. By *problematic*, we mean that the obtained vector of p-values would require a substantial amount of trafficking in order to pass a detection test and this may violate the requirement of a coherent structure for the paths. Typically what is possible is to remove, add, or change particular options or layers, but this is unlikely to alter the distribution exactly in the sought direction. Therefore, if the original p-curve is far from decreasing in the first place, minor adjustments will not suffice and hacking will be complicated.

4.6 Correlation between paths

This subsection seeks to evaluate the theoretical assumption of Proposition 2 that $\rho(p, q) = \rho^{d(p,q)}$ with d from (13). To this purpose, we re-run the paths on subsamples of the original data. We subsample 70% of the data without replacement and run all paths for four variables (*acc*, *mvel1*, *bm_ia* and *mom12m*). We then compute the correlation between all paths across 120 subsamples (results are the same if we only consider 80 or 100 samples).[25] Then, given the distances between the paths, we compute the average correlation between the paths. The resulting patterns are illustrated in Figure 18. Plainly, on average, the assumption that correlation decreases is validated. The shape of the decay seems also to be of power type, with a decay rate of 60%, which is rather strong.

Lastly, we briefly mention the important quantity that is the L^1 norm of the correlation matrix. In our computations, it is of the magnitude of 0.5, which is much too high to approximate the distribution function of the true underlying effect. This number must be reduced, ideally tenfold. There are several ways to achieve this. First, we could select a subset of paths with lower covariance norm, but this is not straightforward to do. Basically, this amounts to pick paths which tell different stories about the effect and remove the ones that are the most redundant. A second option would be to use different sub-samples each time we run the paths. Preliminary tests show that using samples of half the size of the original data reduces the covariance matrix norm by a

[25] On a 2019 iMac and across 5 cores, all 4,608 paths takes 45 minutes to run per sorting variable, hence 100 draws are very time consuming.

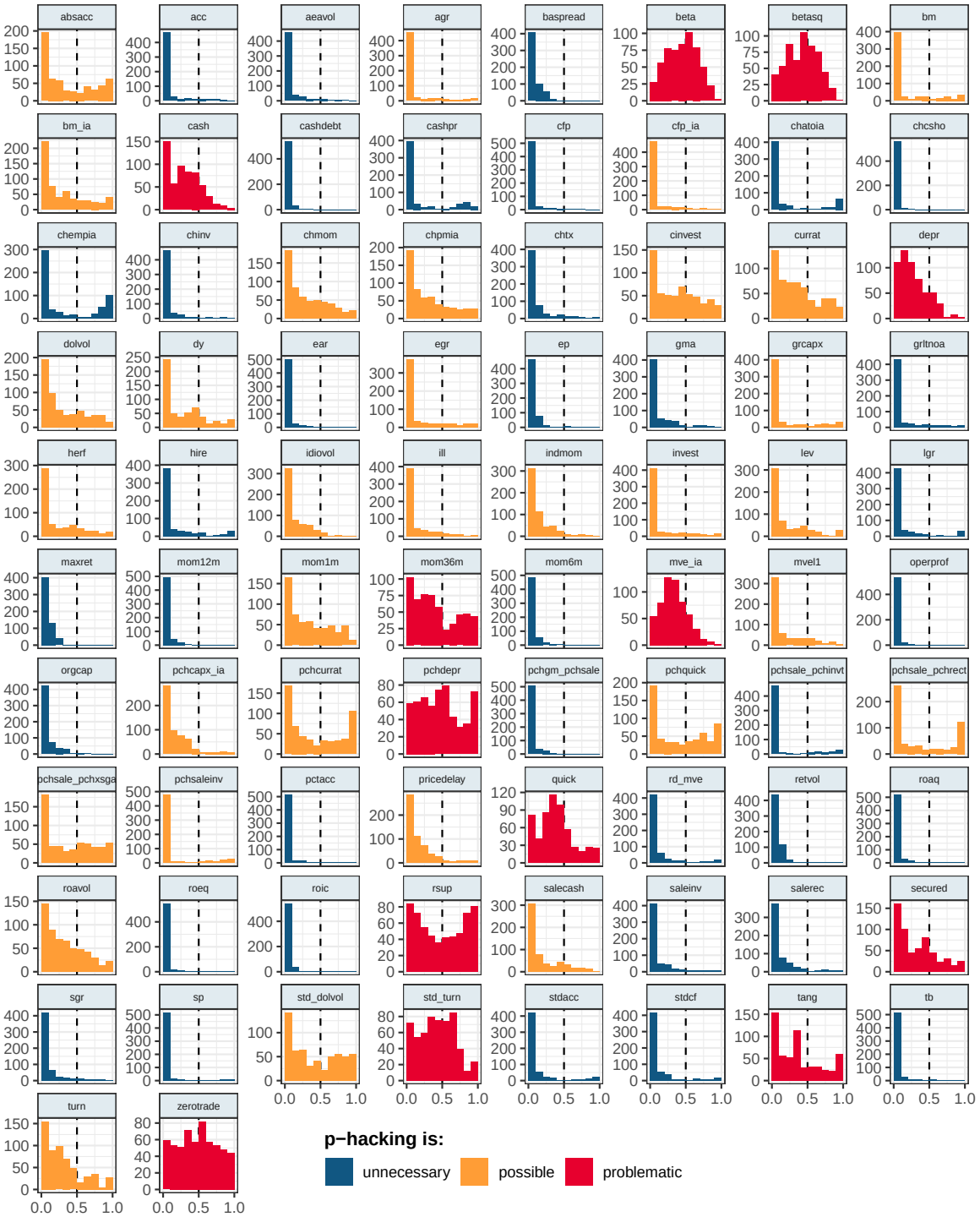

Figure 17: **Distributions of p-values**. We produce the histogram of the p-values, for each sorting variable. There are $I = 10$ breaks, which mark the deciles of the distribution. The vertical lines mark the limit of the critical zone ($x = 1/2$). Colors code the level of κ in Equation (33).

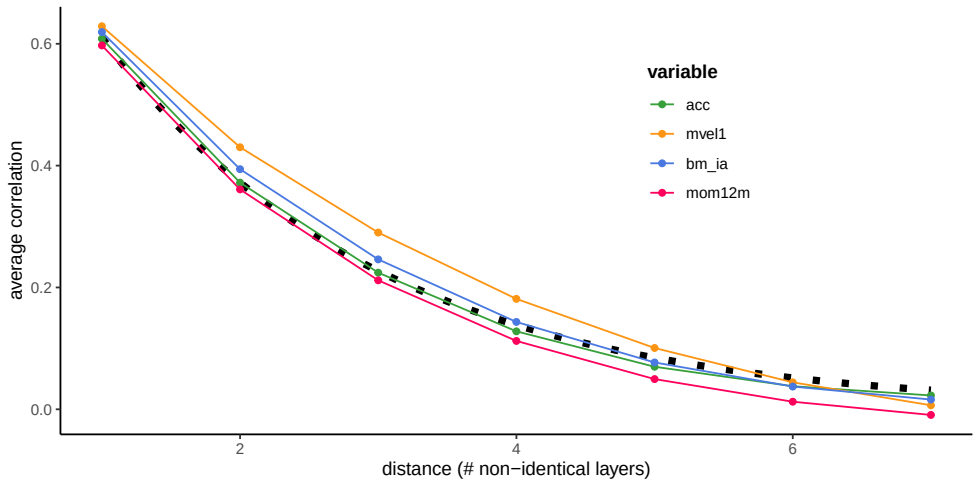

Figure 18: **Correlation across paths**. We depict the decay in average correlation, as function of distances between paths. The black dashed line represents the decay ρ^d for $\rho = 0.61$.

factor 3 to 4. Working with even smaller samples (e.g., 30%) of the sample would shrink the error even further. This is an important matter, but it is left for future research.

5 Application: Fama-MacBeth regressions

5.1 Method and paths

A cornerstone question in financial economics pertains to the evaluation of the risk premium of asset pricing factors. Arguably, the most popular approach to answer this question is the double-pass regression of Fama and MacBeth (1973). For the sake of completeness, we recall the two passes briefly below. First, asset returns are regressed on the target F factors as follows:

$$r_{t,n} = a_n + \sum_{f=1}^{F} b_n^f f_t + e_{t,n}, \quad \text{estimated for all assets, } n. \tag{34}$$

The estimated loadings $\hat{b}_n^f$ are then recycled as independent variable in the second pass in which is run on a date-by-date basis:

$$r_{t,n} = \gamma_{t,0} + \sum_{f=1}^{F} \gamma_t^f \hat{b}_n^f + \epsilon_{t,n}, \quad \text{estimated for all dates, } t. \tag{35}$$

Nevertheless, the methodology leaves several open choices. One important option is how to perform the first pass. For instance, it is possible to estimate the loadings on the full sample, an approach which we henceforth qualify as "*in-sample*", or to estimate them on rolling samples ("*out-of-sample*"). Another important degree of freedom is the set of assets from which these loadings are estimated. For instance, Ang et al. (2020) argue that, contrary to conventional wisdom, using portfolios instead of individual assets "*destroys information by shrinking the dispersion of betas, leading to larger standard errors*".

In Figure 19, we show the diversity of paths for this study, with the following choices and layers:

1. the **factor** for which the risk premium is computed, among the five Fama and French (2015) factors;
2. the base **assets** that are used for the estimation. There are five alternatives: two Fama-French sorted portfolios on book-to-market (25 or 100 portfolios), two Fama-French industry portfolios and 393 individual stocks, available in open source for reproducibility;
3. the **weighting scheme** of portfolios in base assets, either equally-weighted or value-weighted (does not apply to individual stocks);
4. the data **frequency** for the first pass (initial loading estimation), whether it is daily or monthly;
5. the **winsorization level** for returns before the first pass;
6. the **regression type**, i.e., whether the first pass is run on the full sample, or on rolling windows of short or long samples (24 versus 60 months or 120 versus 300 days depending on data frequency).
7. the **winsorization level** for returns before the second pass, which is applied to treat outliers in estimated betas.

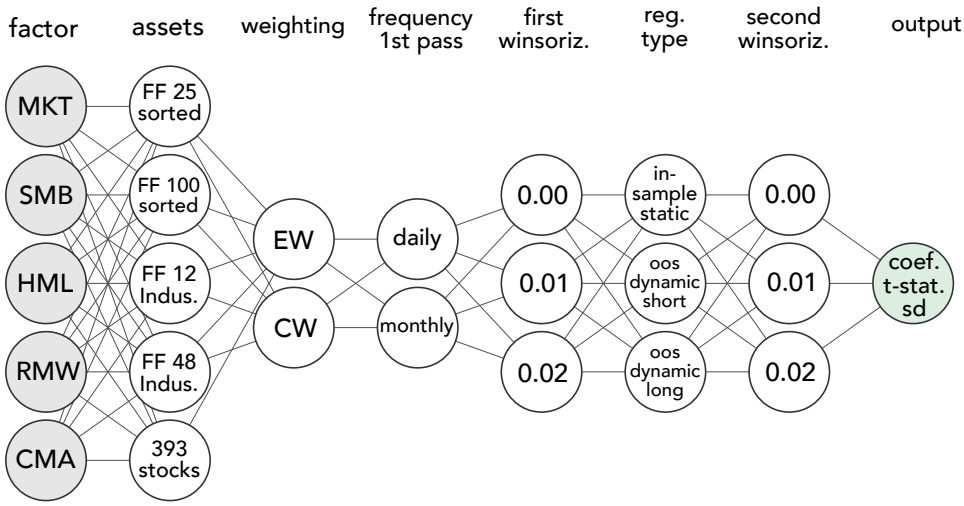

Figure 19: **Forking paths for Fama-MacBeth regressions**. We depict the seven steps of the protocol. Note: for individual stocks, weighting does not apply - it is fixed to EW by default. In total, there are $9 \times 2 \times 3 \times 3 \times 3 = 486$ paths for each factor.

This procedure generates, for each date t and factor f, the $\hat{\gamma}_t^f$, along with the related AIC criteria and standard errors which can be used to compute weights (18) and confidence intervals (20). The results are obtained for the five factors from Fama and French (2015).

5.2 Baseline results

In Figure 20, we plot the average premia for calendar years.[26] First, we compute simple and weighted averages for each month and these values are then averaged (simply) over the year (orange curve). We also report the Bayesian confidence intervals in light blue, but they are so narrow that they are indistinguishable from the corresponding weighted average curves (Equation (22), in dark blue). Results for frequentist averages are qualitatively the same and the intervals remain too thin to discriminate from the means, as in Figure 6 for monthly frequencies. In unreported

[26]The code for this section can be accessed here.

results, we find that simple averages have larger standard errors and thus yield larger hence more conservative intervals. Overall, the different averaging schemes produce quite consistent values, especially for the value factor (HML).

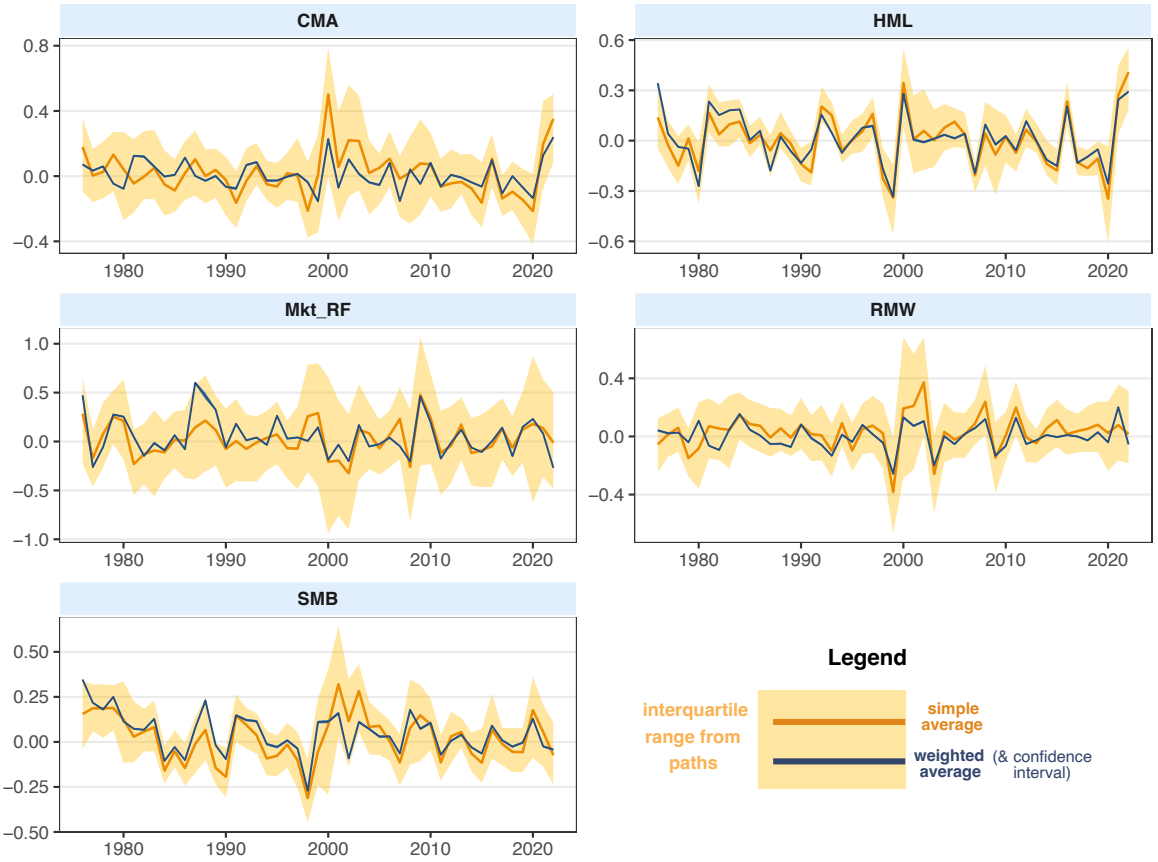

Figure 20: **Fama-MacBeth premia: distribution across time**. We represent some properties of loadings from the second pass of Fama and MacBeth (1973) regressions. Monthly values are averaged at the annual scale to ease readability. The **orange** line shows the simple average of loadings across all 486 paths. The **dark blue** line depicts the Bayesian weighted average from Equation (22). Around these latter averages, the 99% confidence intervals from Equation (20) are indistinguishably thin. Finally, the light yellow shaded areas represent the inter-quartile range across all paths.

Lastly, in the background of the figures, we show the inter-quartile range across all paths. The latter is the widest for the market factor and illustrates the diversity of outcomes and the richness of paths. Notably, we find that the paths often yield values of contrasting signs, which is not surprising since premia oscillate around zero. There are however some instances where a large majority of path values lie either above or below zero: they correspond to cases when many paths agree on the sign of the premium (e.g., in 1992 or 2021 for HML or in 1990 for SMB).

5.3 Which design options matter?

To illustrate the conditional averages proposed in Section 2.4.2, we focus on the layers for which there are 3 possible options in our protocol: they are the first pass regression type and the two winsorization levels (before and after the first pass).

Our results are gathered in Table 3. The left part of the table reports the average premium of factors when fixing either of the three options o_j. For instance, for Panel B and C, o_1 corresponds to an absence of winsorization, while o_2 and o_3 winsorize at the 1% and 2% level, respectively. For panel A, the o_j pertain to the type of first pass regression (sampling scheme). The right part of the table gathers the difference between the options and the statistical significance of the related simple t-test. We see that for panel C, differences are hardly significant, which means that the second winsorization, after the loadings have been estimated, has little impact on premia estimates.

Factor	Means of options o_j			Difference in means		
	o_1	o_2	o_3	$o_1 - o_2$	$o_1 - o_3$	$o_2 - o_3$
Panel A: Regression type (first pass)						
CMA	0.149	0.187	0.264	-0.038 *	-0.115 ***	-0.077 **
HML	0.091	0.110	-0.115	-0.019	0.206 ***	0.225 ***
MKT	0.141	0.147	0.267	-0.006	-0.126 **	-0.12 **
RMW	0.135	0.173	0.418	-0.038 *	-0.283 ***	-0.245 ***
SMB	0.180	0.283	0.068	-0.103 ***	0.112 ***	0.215 ***
Panel B: Winsorization threshold before the first pass						
CMA	0.430	0.107	0.070	0.323 ***	0.36 ***	0.037
HML	-0.033	0.038	0.066	-0.071 **	-0.099 ***	-0.028
MKT	0.074	0.189	0.298	-0.115 **	-0.224 ***	-0.109 *
RMW	0.194	0.267	0.283	-0.073 **	-0.089 ***	-0.016
SMB	0.202	0.146	0.173	0.056 *	0.029	-0.027
Panel C: Winsorization threshold after the first pass						
CMA	0.210	0.204	0.193	0.006	0.017	0.011
HML	0.021	0.023	0.027	-0.002	-0.006	-0.004
MKT	0.150	0.194	0.223	-0.044	-0.073	-0.029
RMW	0.235	0.253	0.256	-0.018	-0.021	-0.003
SMB	0.164	0.177	0.180	-0.013	-0.016	-0.003

Table 3: **Which options matter?** We compute the average premia for choices with three options o_1, o_2 and o_3. These choices are short rolling regression (o_1), long dynamic rolling regression (o_2) and static regression (o_3) for the first pass and 0.00, 0.01 and 0.02 respectively for the winsorization thresholds, before and after the first pass. The significance levels are: (***)<0.001<(**)<0.01<(*)<0.05.

In contrast, in the first two panels, we report some substantial differences. From panel B, we infer that switching from no winsorization to 1% or 2% does alter the results, though not always in the same direction, depending on the factor. The figures from panel A imply that the choice of samples for the first pass also matters. The largest changes occur between full sample versus both types of rolling samples. Interestingly, switching from small to longer dynamic samples always has a negative impact on premia, though we cannot rationalize why that may be the case.

5.4 Comparison with prior work

We wish to compare the distribution of the 486 paths we generated for each date with the values of prior studies, namely Fama and MacBeth (1973) and Ang et al. (2020). For simplicity, we will only compare premia that are averaged over the longest periods in these articles.

In Fama and MacBeth (1973), Table 4 in Section V-B, the average γ value for the market premium is 0.0085 for the longest sample, from 1935 to 1968. Ang et al. (2020), in Table 5 in Section

IV-A, report four values, depending on the set of assets used for the estimation: 0.0114, 0.0158, 0.0173 and 0.0479. We locate these values in comparisons to the ones we obtained with vertical lines in Figure 21.

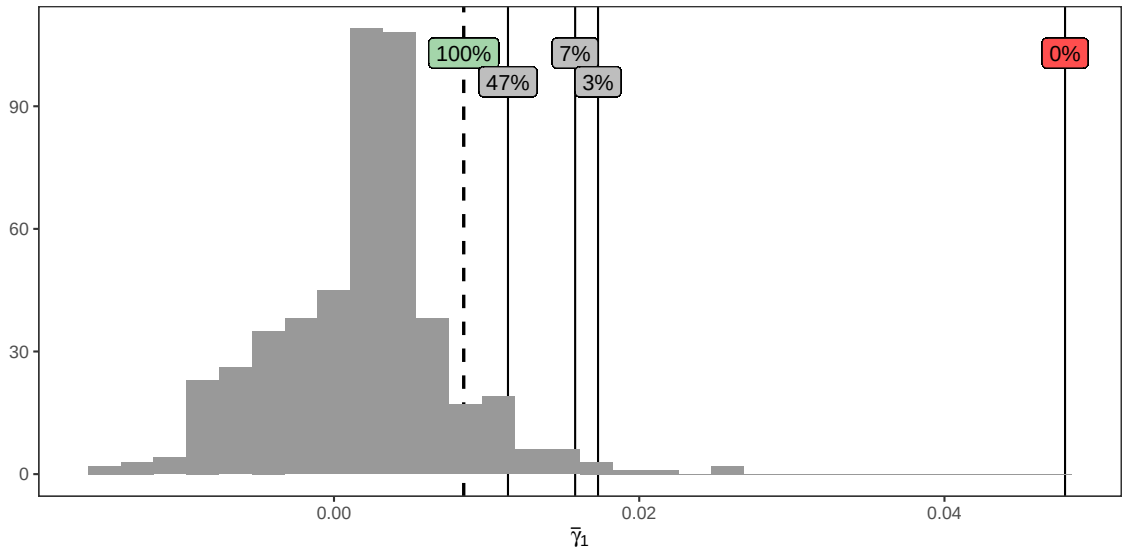

Figure 21: **Comparison with prior work (market factor).** We depict the distribution of average monthly premia of the market factor spanned by the 486 paths. In addition, we locate the corresponding values for the market factor from the studies of Fama and MacBeth (1973) and Ang et al. (2020) with dashed and full vertical lines, respectively. The colored rounded squares indicate the ease to confirm (EtC) of the reported value with respect to the Gaussian distribution fitted on the grey histogram (mean and standard deviation). The EtC is equal to one minus (30) with $q = 0.9$. The colors code the facility to reproduce the original outcome.

In addition, we specify the ease-to-replicate indicator at the top of the vertical lines. We find that the value reported by Fama and MacBeth (1973) is very reasonable, as it lies not too much to the right of the distribution obtained from the paths. With regard to the average premia found in Ang et al. (2020), the ease to reproduce results is more mixed and lukewarm, with one value (4.8% per month) which is found to be virtually impossible to reach with the paths that we were able to span. This large value correspond to the estimation with the individual stocks, which seem to suggest that granular test assets yield higher estimated premia. We are able to test this conjecture and do not find much support for it. In Figure 25 in the Appendix, we plot average absolute premia across all factors and test asset groups and do not find a strong relationship between the granularity of assets and the magnitude of premia. If anything, it is the least granular assets (the 12 industries) that are associated with the largest absolute premia.

A last word on sample sizes. In Fama and MacBeth (1973), the average premia for the market change substantially from period to period. For example, it is equal to 0.163 in 1935-1945 and to 0.143 in 1961-1968. These values are much larger than the estimate over the total period and would obtain an EtC score well below 50%. Hence, again, considering chronologically deep samples increases the odds of confirmation.

6 Conclusion

Amid debates in some scientific communities about the validity of empirical results, we make the case for an exhaustive approach. Formally, we suggest to report results for a large number of

design choices, as if extensive robustness checks were in fact constituent of the baseline research protocol. Small variations in designs allow the generation of many estimates and test statistics. The distribution of these statistics can help figure out if one configuration yielded a favorable outlier, or if the sought effect is indeed statistically strong. Moreover, having many coefficient or statistics at one's disposal allows to resort to aggregation so as to obtain more robust estimates and confidence intervals.

The application of these ideas to three exercises in financial economics confirms the intuition that by considering many design options, the range of outcomes can increase rapidly. Because of this, it is important to report which particular choices in the protocol may shift the distribution of outcomes. Moreover, spanning large numbers of paths allow to determine when prior results are plausible and robust - or not. We craft an indicator called the *ease-to-corroborate* (EtC) which evaluates how realistic effect sizes published in prior studies can be. We find substantial heterogeneity in the asset pricing literature with respect to this criterion, with some contributions that report very reproducible outcomes and others that propose results that appear much harder to replicate.

Forking paths can also be very useful to generate conclusions that are more trustworthy, especially for investment purposes. Averaging returns, premia or loadings across many configurations strengthens inference. We find high cross-period correlation between anomalies returns' once they have been averaged across many paths. This removes the risk of an outlier point from one specific set of implementation choices and increases the odds of generalization out-of-sample.

Paths can also be applied to multiple testing. They allow to generate outcomes that are less homogeneous, compared to bootstrapped series. This produces distributions for maximum statistics which have heavier tails. Consequently, the corresponding significance thresholds are higher. For asset anomalies to hold, our framework requires that they remain statistically profitable under various weighting schemes and across several sub-periods. Our second empirical analysis finds that the bar for t-statistics in portfolio sorts should be raised to 8.2, an almost prohibitive level that is much higher than those typically used in the literature. However, higher decision hurdles also come at the cost of more false negatives, which may or may not matter, especially for investment purposes.

There are of course several limitations to our suggestions. First, it is possible to push the limits of data-snooping to the extreme by reporting only the combinations of design choices that fit a particular narrative, but this is arguably cumbersome. Second, given the amount of time required to generate comprehensive results, the research question must be inherently simple. Each path should not take more than a handful of minutes, so that hundreds, or thousands, of them can be generated in less than one day. The aim of the paper is clearly not to increase the carbon footprint of researchers. Long computation times may contribute to this footprint and we refer to Mariette et al. (2021) for a discussion on this matter. This is why a precise framing of the research question, as well as its relevant ramifications, is imperative to avoid superfluous digressions.

A Examples of Lipschitz constants

This section relies heavily on norms. For vectors, we work with L^p norms: $\|\boldsymbol{v}\|_p^p = \sum_{n=1}^N |v_n|^p$ and for $(N \times K)$ matrices we will consider the following:

- $\|\boldsymbol{M}\|_2^2 = \sum_{n=1}^N \sum_{k=1}^K M_{n,k}^2$: Frobenius Norm;
- $\|\boldsymbol{M}\|_1 = \max_k \sum_{n=1}^N |M_{n,k}|$: maximum absolute column sum;
- $\|\boldsymbol{M}\|_\infty = \max_n \sum_{k=1}^K |M_{n,k}|$: maximum absolute column row.

A.1 Descriptive statistics

Sample moments and other mainstream metrics play an important in empirical studies. We begin our journey of illustrations with the simplest of them all: the sample mean. In the sequel, we will use the notation f as generic mapping. The case of the (biased) **sample variance** is more tricky:

$$\begin{aligned}
\|f(\mathbb{d}) - f(\boldsymbol{d})\|_p &= \left| \frac{1}{N}\sum_{n=1}^N \left(\mathbb{d}_n - \frac{1}{N}\sum_{n=1}^N \mathbb{d}_n\right)^2 - \frac{1}{N}\sum_{n=1}^N \left(d_n - \frac{1}{N}\sum_{n=1}^N d_n\right)^2 \right| \\
&= \frac{1}{N}\left| N(\bar{d}^2 - \bar{\mathbb{d}}^2) + \sum_{n=1}^N \mathbb{d}_n^2 - d_n^2 \right| \\
&= \frac{1}{N}\left| N(\bar{d} - \bar{\mathbb{d}})(\bar{d} + \bar{\mathbb{d}}) + \sum_{n=1}^N (\mathbb{d}_n + d_n)(\mathbb{d}_n - d_n) \right| \\
&\leq |\bar{\mathbb{d}} + \bar{d}| \times |\bar{\mathbb{d}} - \bar{d}| + \frac{d^*}{N}\left| \sum_{n=1}^N (\mathbb{d}_n - d_n) \right| \\
&\leq c_p \|\mathbb{d} - \boldsymbol{d}\|_p, \qquad (36)
\end{aligned}$$

where $c_p = N^{-1/p}(|\bar{\mathbb{d}} + \bar{d}| + d^*)$, with $d^* = \max_n |\mathbb{d}_n + d_n|$. $\bar{\mathbb{d}}$ and $\bar{d}$ are the sample means. The last inequality comes from (7). In this case, and as will be recurrent, the constant depends on the magnitude of the series. To remove this dependence, it is imperative to specify some properties of the vectors (e.g., if they belong to the unit sphere, or if their range is restricted to particular intervals). This comment holds for the remainder of the paper, as many constants will be input-dependent below.

Typically, for the sample **covariance**, we have that

$$\begin{aligned}
|f(\mathbb{d}_1, \boldsymbol{d}_1) - f(\mathbb{d}_2, \boldsymbol{d}_2)| &= |(\mathbb{d}_1 - \bar{\mathbb{d}}_1)'(\boldsymbol{d}_1 - \bar{\boldsymbol{d}}_1) - (\mathbb{d}_2 - \bar{\mathbb{d}}_2)'(\boldsymbol{d}_2 - \bar{\boldsymbol{d}}_2)| \\
&= |(\mathbb{d}_1 - \bar{\mathbb{d}}_1 - (\mathbb{d}_2 - \bar{\mathbb{d}}_2))'(\boldsymbol{d}_1 - \bar{\boldsymbol{d}}_1) - (\mathbb{d}_2 - \bar{\mathbb{d}}_2)'(\boldsymbol{d}_2 - \bar{\boldsymbol{d}}_2 - (\boldsymbol{d}_1 - \bar{\boldsymbol{d}}_1))| \\
&\leq \frac{\|\boldsymbol{d}_1 - \bar{\boldsymbol{d}}_1\|_1}{N}(|\mathbb{d}_1 - \mathbb{d}_2| + |\bar{\mathbb{d}}_1 - \bar{\mathbb{d}}_2|) + \frac{\|\mathbb{d}_2 - \bar{\mathbb{d}}_2\|_1}{N}(|\boldsymbol{d}_1 - \boldsymbol{d}_2| + |\bar{\boldsymbol{d}}_1 - \bar{\boldsymbol{d}}_1|)
\end{aligned}$$

which is again a similar form.

Let us now mention the **maximum** of vectors. We have that

$$\max_n \mathbb{d}_n = \max_n [\mathbb{d}_n - d_n + d_n] \leq \max_n |\mathbb{d}_n - d_n| + \max_n d_n,$$

so that

$$\|\max_n \mathbb{d}_n - \max_n d_n\|_\infty = |\max_n \mathbb{d}_n - \max_n d_n| \leq \max_n |\mathbb{d}_n - d_n| = \|\mathbb{d} - \boldsymbol{d}\|_\infty,$$

i.e, the Lipschitz constant in this case is one. Straightforwardly, the same applies to the minimum operator.

A.2 Other examples of Lipschitz constants

In empirical studies, the **data collection** stage is the hardest to model, because of its heterogeneity. It can be quite constrained if data comes from a provider (e.g., WRDS, Bloomberg, etc.), in which case the researcher has a few degrees of freedom: which variables to import, for which universe, at which frequency, over which time frame, etc. Providers often update their data so that downloading a sample at two different periods may generate discrepancies if series are not kept point-in-time. This has been recently shown for the Fama-French factors in Akey et al. (2023). There is also relatively little room for initiative in economics or physics when working with official series, such as GDP output, inflation, unemployment, CO_2 concentration, temperatures, etc.

However, when the study is based on surveys, the researcher has more latitude, and we for instance refer to the guide of Bergman et al. (2020) for an overview of the range of options in that case. In qualitative studies, there is also an important coding phase (see, e.g., the review by Basit (2003)), which is difficult to model neatly and efficiently.

For all these reasons, we commence this section with the step that comes right *after* data collection, namely **data cleaning**. We underline that Mitton (2022) reports that "*the methodological decisions that affect statistical significance the most are dependent variable selection, variable transformation, and outlier treatment*". In this subsection, we tackle all of these elements. Lastly, the purpose of the section is to show that most classical operations on data can be represented as Lipschitz mappings. It is not to provide sharp constants for these mappings.

A.2.1 Data cleaning

The first issue that most, if not all, researchers encounter, is **missing data**. When working with time-series, a common practice is to impute with the most recent well-defined point prior to the missing value, if it exists. Another option is to resort to cross-sectional means or medians. Interpolation is usually avoided because it introduces a forward-looking bias. In this subsection, to ease the exposition, we make strong assumptions on the vectors on which the imputation mapping will operate.

Formally, we consider two vectors $\mathbb{d}$ and $\boldsymbol{d}$ such that S is the *common* set of indices for which a value is missing. The fact that S is common to the two vectors comes from the fact that we need $\|\mathbb{d} - \boldsymbol{d}\|$ to be well defined. In the above norm, two points that are not defined are assumed to be equal, but if one value is defined and the other is not, there is no unambiguous way to proceed. For simplicity, we assume that 1 does not belong to the S set, so that the imputation values will always be defined. In addition, we impose that the indices in S are never consecutive numbers, though this assumption can be relaxed easily. We then have

$$\|f(\mathbb{d}) - f(\boldsymbol{d})\|_p^p = \sum_{n\in S} |\mathbb{d}_{n-1} - d_{n-1}|^p + \sum_{n\notin S} |\mathbb{d}_n - d_n|^p \leq 2\|\mathbb{d} - \boldsymbol{d}\|_p^p$$

where the last inequality comes from the fact that the values that precede missing points get counted twice. The Lipschitz constant is not very sharp in this case.

Other examples of methods include cross-sectional imputation, whereby a missing value is replaced by the cross-sectional mean (or median) across other observations. In this case, via inequality (21) it is also possible to derive a Lipshitz constant for mean-driven imputation. Parametric imputation based on some distributional assumption follow the same logic, though their treatment is substantially more involved.

One extreme solution when facing missing data is simply the **removal of observations**. To illustrate this issue, we consider two matrices of numerical data $\mathbb{D}$, $\boldsymbol{D}$ with equal sizes, N rows and M columns. In line with the above assumptions, we write S for the (common) indices of their rows which contain missing points. Naturally, we again assume that the cardinal of S is

much smaller than the total number of rows N. In this case, it is straightforward that for the usual matrix norms (Frobenius, $\|\cdot\|_1$ and $\|\cdot\|_\infty$), the Lipschitz constant is one at most, i.e., that

$$\|f(\mathbb{D}) - f(\boldsymbol{D})\| \leq \|\mathbb{D} - \boldsymbol{D}\|.$$

Similarly, the researcher may want to remove columns (instead of rows) of the data because of **co-linearity** issues. The Lipschitz constant of such an operation is also one at most.

Another important stage in data processing is **outlier management**. One of the most frequently used tools to this purpose is **winsorization**, whereby extreme values are replaced by given quantiles, often at the 1% and 99% levels.

Without loss of generality, let us assume that the vector $\mathbb{d}$ is ordered, i.e., that $\mathbb{d}_1 < \cdots < \mathbb{d}_N$. For a given integer $k \ll N/2$, the winsorization operator is defined as:

$$f(\mathbb{d}_n) = \begin{cases} \mathbb{d}_n & \text{if } n \in [k+1, N-k] \\ \mathbb{d}_{k+1} & \text{if } n \leq k \\ \mathbb{d}_{N-k} & \text{if } n > N-k \end{cases}, \tag{37}$$

so that exactly $2k$ values are replaced: the most extreme k values in both tails. If we abusively write $f(\mathbb{d})$ for the vector of $f(\mathbb{d}_n)$ values, it holds that

$$\begin{aligned} \|f(\mathbb{d}) - f(\boldsymbol{d})\|_1 &= \sum_{n=1}^{k} |\mathbb{d}_{k+1} - \mathbb{d}_{k+1}| + \sum_{n=k+1}^{N-K} |\mathbb{d}_n - \mathbb{d}_n| + \sum_{n=N-K+1}^{N} |\mathbb{d}_{N-k} - \mathbb{d}_{N-k}| \\ &= K(|\mathbb{d}_{k+1} - \mathbb{d}_{k+1}| + |\mathbb{d}_{N-k} - \mathbb{d}_{N-k}|) + \sum_{n=k+1}^{N-K} |\mathbb{d}_n - \mathbb{d}_n| \\ &\leq (1+K)\|\mathbb{d} - \boldsymbol{d}\|, \end{aligned}$$

where the constant is clearly sub-optimal. It could be improved but would then rely on the properties of the underlying vectors.

A.2.2 Variable engineering

Once the data has been cleaned, the researcher will often perform additional adjustments. We list a few below.

Normalization is a common step in data preparation: it ensures that all variables have roughly the same scales. This is convenient when one wants to compare effect sizes for example. There are several ways to proceed, such as standardization, or min-max rescaling. Let us analyze the former:

$$\begin{aligned} \|f(\mathbb{d}) - f(\mathbb{d})\|_2 &= \left\| \frac{\mathbb{d} - \boldsymbol{m}_{\mathbb{d}}}{\sigma_{\mathbb{d}}} - \frac{\mathbb{d} - \boldsymbol{m}_{\mathbb{d}}}{\sigma_{\mathbb{d}}} \right\|_2 = \sigma_{\mathbb{d}}^{-1}\sigma_{\mathbb{d}}^{-1}\|\sigma_{\mathbb{d}}(\mathbb{d} - \boldsymbol{m}_{\mathbb{d}}) - \sigma_x(\mathbb{d} - \boldsymbol{m}_{\mathbb{d}})\|_2 \\ &= \sigma_{\mathbb{d}}^{-1}\sigma_{\mathbb{d}}^{-1}\|(\sigma_{\mathbb{d}} - \sigma_{\mathbb{d}} + \sigma_{\mathbb{d}})(\mathbb{d} - \boldsymbol{m}_{\mathbb{d}}) - \sigma_{\mathbb{d}}(\mathbb{d} - \mathbb{d} + \mathbb{d} - \boldsymbol{m}_{)}\|_2 \\ &\leq \sigma_{\mathbb{d}}^{-1}\sigma_{\mathbb{d}}^{-1}|\sigma_{\mathbb{d}} - \sigma_{\mathbb{d}}| \times \|\mathbb{d} - \boldsymbol{m}_{\mathbb{d}}\|_2 + \sigma_{\mathbb{d}}^{-1}\|\mathbb{d} - \mathbb{d}\|_2 + \sigma_{\mathbb{d}}^{-1}\|\boldsymbol{m}_{\mathbb{d}} - \boldsymbol{m}_{\mathbb{d}}\|_2 \qquad (38) \\ &\leq \sigma_{\mathbb{d}}^{-1}\sqrt{N}\frac{|\sigma_x^2 - \sigma_{\mathbb{d}}^2|}{\sigma_{\mathbb{d}} + \sigma_{\mathbb{d}}} + \sigma_{\mathbb{d}}^{-1}\|\mathbb{d} - \mathbb{d}\|_2 + \sigma_{\mathbb{d}}^{-1}N^{-1/2}\|\mathbb{d} - \mathbb{d}\|_2 \qquad (39) \\ &\leq c\|\mathbb{d} - \mathbb{d}\|_2 \end{aligned}$$

where $\boldsymbol{m}_{\mathbb{d}}$ is the constant mean vector of $\mathbb{d}$, $\sigma_{\mathbb{d}}$ its standard deviation and

$$c = \sigma_{\mathbb{d}}^{-1}(1 + N^{-1/2} + \sqrt{N}c_1(\sigma_{\mathbb{d}} + \sigma_{\mathbb{d}})^{-1}),$$

the constant c_1 being the one from Inequation (36) for $p = 1$. We have used that $\|\mathbb{d} - \boldsymbol{m}_{\mathbb{d}}\|_2 = \sqrt{N}\sigma_{\mathbb{d}}$ and applied (7) and (36) in lines (38) and (39), respectively.

Sometimes, when working with time-series, the model requires stationary variables, but the collected data is integrated and has unit roots. In other contexts, the level of the independent variable may matter less than its variations from a predictive standpoints. Thus it is relevant to consider variable differences in these settings too. In any case, the solution is **differentiation**:

$$f(\mathbb{d}_n) = \begin{cases} \text{NA} & \text{if } n = 1 \\ \mathbb{d}_n - \mathbb{d}_{n-1} & \text{otherwise} \end{cases} . \tag{40}$$

In practice, the first missing point is often removed so that the resulting vector has length $N - 1$. For two numerical vectors with no missing points $\mathbb{d}$ and d,

$$\begin{aligned} \|f(\mathbb{d}) - f(\mathbb{d})\|_1 &= \sum_{n=2}^{N} |\mathbb{d}_n - \mathbb{d}_{n-1} - \mathbb{d}_n + \mathbb{d}_{n-1}| \\ &\leq \sum_{n=2}^{N} |\mathbb{d}_n - \mathbb{d}_n| + |\mathbb{d}_{n-1} - \mathbb{d}_{n-1}| \leq 2\|\mathbb{d} - \mathbb{d}\|_1 \end{aligned}$$

It may also happen that researchers seek to explain long term effects. For instance, in the predictability literature, there is a debate between short-term and long-term predictability. At a first order approximation, long term returns can be viewed as **cumulative sums** of shorter horizon returns, which is why we briefly mention the topic below. For a given well-defined numerical vector, we have in this case, for $n > 0$,

$$f(\mathbb{d}_n) = \sum_{k=1}^{n} \mathbb{d}_k,$$

and

$$\|f(\mathbb{d}) - f(\boldsymbol{d})\|_1 = \sum_{n=1}^{N} \left| \sum_{k=1}^{n} \mathbb{d}_k - \sum_{k=1}^{n} d_k \right| \leq \sum_{n=1}^{N} \sum_{k=1}^{n} |\mathbb{d}_k - d_k| \leq N\|\mathbb{d} - \boldsymbol{d}\|_1,$$

where the bound may seem loose, but can be sharp if $\mathbb{d}_n = d_n = 0$ for $n > 1$ for instance.

To conclude this subsection, we acknowledge that many more operations exist in the data preparation phase and some would require a lengthy treatment. For instance, joining procedures that merge two tables according to a common key are a widespread practice. They are however more complex to handle and we leave their analysis to future work.

A.2.3 Testing

An ubiquitous tool in the researcher's arsenal is the **linear regression**. Given a matrix of independent variables $\boldsymbol{X}$ and the vector of dependent variable $\boldsymbol{y}$, the ordinary least square (OLS) estimator $\boldsymbol{b}$ that minimizes the quadratic error

$$e^2(\boldsymbol{X}, \boldsymbol{y}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) \tag{41}$$

is

$$\boldsymbol{b}(\boldsymbol{X}, \boldsymbol{y}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \tag{42}$$

where $\boldsymbol{v}'$ denotes the transpose of $\boldsymbol{v}$ (vector or matrix). In the sequel, we will always assume that the inverse matrix is well-defined. The issue is then that there are two inputs. Given one of them,

it is possible to intuitively deduce data-specific Lipschitz constants. For instance, if $\boldsymbol{X}$ is fixed, then factorizing the $\boldsymbol{X}$-dependent matrices yields

$$\|\boldsymbol{b}(\boldsymbol{X},\boldsymbol{y})-\boldsymbol{b}(\boldsymbol{X},\boldsymbol{z})\| \leq c_X\|\boldsymbol{y}-\boldsymbol{z}\|. \tag{43}$$

The case when $\boldsymbol{y}$ is fixed is less straightforward but can be handled with suitable norms. The general case when both $\boldsymbol{X}$ and $\boldsymbol{y}$ are subject to perturbation is more intricate. It is reviewed in Section 5 of Grcar (2003). One foundational result (Golub and Wilkinson (1966)) is that if $\|\boldsymbol{X}\|_2 = \|\boldsymbol{y}\|_2 = 1$, then

$$\|\boldsymbol{b}(\boldsymbol{X},\boldsymbol{y})-\boldsymbol{b}(\boldsymbol{Z},\boldsymbol{v})\|_2 \leq c(\|\boldsymbol{X}-\boldsymbol{Z}\|_2+\|\boldsymbol{y}-\boldsymbol{v}\|_2)+R, \tag{44}$$

where c depends on the smallest singular value of $\boldsymbol{X}$, on $\|\boldsymbol{b}\|_2$, and on the quadratic error e^2 defined in Equation (42). The matrix norms are of Frobenius type: $\|\boldsymbol{X}\|_2^2 = \text{tr}(\boldsymbol{X}\boldsymbol{X}')$, where $\text{tr}(\cdot)$ is the trace operator. The above result holds in the case when the norms are arbitrarily small and the residual term R is a second order term which is quadratic in the maximum of the two norms.

While the coefficients in linear regressions (or more general models) are undoubtedly analyzed by researchers, it is their **statistical significance** which often matters more because it will determine if the effect revealed by the study is strong enough.

In the case of a linear model, the expressions for the t-statistics are $t_k = b_k/\sqrt{s^2 S_k}$, where S_k is the k^{th} diagonal element of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ and $s^2 = e^2/(N-K)$, with K being the number of columns of $\boldsymbol{X}$ (Equation 4-47 in Greene (2018)). Lipschitz numbers can be obtained for S_k (see e.g., Demmel (1992) and Loh and Tan (2018)) and for $s^2 = (N-K)^{-1}(\boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y})$ as well. From these numbers, it is possible to derive a Lipschitz constant similar to (44) for the statistic, based on norms that depend on the inputs.

Often, models are tested with several control variables and model permutations, in order to check the robustness of the initial specification. It is even possible to aggregate coefficients of multiple specifications via weighting, as in Hansen (2007) and Zhang and Liu (2019). In linear models, adding columns to the matrix $\boldsymbol{X}$ is often handled via the Frish-Waugh-Lovell theorem. The Lipschitz continuity of this operation with respect to estimates and t-statistics is currently out of the scope of the paper (it requires to replace the concatenation $[\boldsymbol{X}_1\ \boldsymbol{V}_1]$ with $[\boldsymbol{X}_2\ \boldsymbol{V}_2]$ as the independent matrix in Equation (42)). In the field of randomized experiments, Muralidharan et al. (2022) analyze the change in significance for coefficients of short versus long models.

Naturally, modern studies rely on much more complex apparatus, including structural equations, difference-in-differences, dynamic panels, HAC estimators (White (1980), Newey and West (1987) and Andrews and Monahan (1992), to cite the most commonly used), etc. Given the exhaustiveness and complexity of the related methods, we cannot treat them comprehensively, but it is possible that many of them can be described as Lipschitz operators. Obviously, the variety of strata in the design of empirical studies is such that most intermediate steps cannot be listed in the present paper - one reason being that many of them can be discipline-specific.

B Handling the correlations between paths

Intuitively, close paths are expected to yield a vicinity of outcomes. Given two paths $p = (r_{p,1}, \dots, r_{p,J})$ and $q = (r_{q,1}, \dots r_{q,J})$, we recall the assumption from Equation (13): $d(p,q) = \#\{j, r_{p,j} \neq r_{q,j}\}$. This assumes that all layers have the same importance, and the rationale for this simplification is that it will greatly help the analytical derivations below.

Below, we prove Proposition 2.

The main issue lies with the distribution of the distance measures $d(p,q)$. They are supported on the set of integers between zero (when $p = q$) and J, which occurs when the paths have no option in common. In order to characterize the number of paths which have exactly a distance of d, we introduce the notion of elementary symmetric polynomials:

$$e_k(x_1, \dots, x_J) = \sum_{1 \leq j_1 < \dots < j_k \leq J} x_{j_1} \dots x_{j_k}, \tag{45}$$

in which there are $\binom{J}{k}$ terms. Let us start with the simplest case, $d = 1$ ($d = 0$ being trivial). We are seeking all the pairs of paths with a distance of one. If we pick one path, say, the first one, then we first need to find which mapping will be the one where the difference occurs, and then count how many options there are. In this case, there are J possible mappings and for each mapping j, the number of options is $r_j - 1$, that is, all options, except the one from the first path. The number of possibilities is thus $\sum_{j=1}^{J}(r_j - 1)$.

More generally, or any path p, the number of other paths which have an arbitrary distance of d (with p) is

$$\sum_{q=1}^{\binom{J}{d}} \prod_{s=1}^{d} (r_{r_{q,s}} - 1) = e_d(r_1 - 1, \dots, r_J - 1), \quad d \geq 1 \tag{46}$$

where the sum is over all permutations of relevant mappings and the product counts the number of remaining options. Note that the number does not depend on p. If we then sum over $d = 0, 1, \dots, J$, we recover the total number of paths $\prod_{j=1}^{J} r_j$ via Vieta's formula:

$$\prod_{i=1}^{n}(x - z_i) = x^n + \sum_{k=1}^{n}(-1)^k e_k(z_1, \dots, z_n) x^{n-1}, \tag{47}$$

which we have combined, for $x = 1$, with $\prod_{j=1}^{J} r_j = \prod_{j=1}^{J}(1 - (1 - r_j))$ and

$$e_k(r_1 - 1, \dots, r_J - 1) = (-1)^k e_k(1 - r_1, \dots, 1 - r_k).$$

Importantly, one term in the r.h.s. of (47) is isolated and corresponds to the case $d = 0$, so that we do not forget to count the initial path p.

Coming back to the norm of the correlation matrix, the fact that the distribution of distances does not depend on the paths allows us to consider only one sum (multiplied P times), as follows:

$$\|\mathbf{\Sigma}_P\|_1 = P^{-1}\left(1 + \sum_{p=2}^{P} \rho^{d(1,p)}\right) = P^{-1}\left(1 + \sum_{d=1}^{J} \rho^d \times e_d(r_1 - 1, \dots, r_J - 1)\right).$$

In the second equality, we switch from a given ordering of the paths to the sum of correlation values (ρ^d) multiplied by the number of times they appear in the sum - see Equation (46). Given the fact that each term in the polynomial is of order d, we can factor in the ρ:

$$
\begin{aligned}
\|\mathbf{\Sigma}_P\|_1 &= P^{-1}\left(1+\sum_{d=1}^{J} e_d(\rho(r_1-1),\dots,\rho(r_J-1))\right)\\
&= P^{-1}\left(1+\sum_{d=1}^{J}(-1)^k e_d(\rho(1-r_1),\dots,\rho(1-r_J))\right)\\
&= P^{-1}\left(\prod_{j=1}^{J}(1+\rho(r_j-1))\right) \quad \text{via (47) for } \; x=1\\
&= \prod_{j=1}^{J}\frac{1+\rho(r_j-1)}{r_j}, \quad \text{because } \; P=\prod_{j=1}^{J} r_j.
\end{aligned}
$$

We recall that $r_j \geq 2$, so that $\frac{1+\rho(r_j-1)}{r_j} < 1$ for $\rho < 1$. Hence, as $J \to \infty$, $\|\mathbf{\Sigma}_P\|_1 \to 0$. However, if J is fixed and $r_j \to \infty$, then $\|\mathbf{\Sigma}_P\|_1$ does not shrink to zero - unless $\rho = 0$, which is a trivial case we exclude.

C Discussion on p-hacking tests

With forking paths, we should not be interested in p-hacking tests such as the ones in Elliott et al. (2022) because the purpose of exhaustive paths is precisely to *avoid* p-hacking in the first place. However, out of curiosity and in full disclosure, we have used these tests on the series of paths obtained from our two empirical studies and got disappointing results: while we did not hack, the tests in some cases concluded that we did. For instance, testing for p-hacking for the p-values with red histograms in Figure 17 will clearly lead to reject the null of no p-hacking.

Let us understand why that may be the case. In Elliott et al. (2022), there are 3 important distributions: F_h, the one of the test statistic, F the distribution that determines the critical values under the null, and Π, the true distribution of the effect under scrutiny. The latter does not matter, as all tests in Elliott et al. (2022) hold irrespective of Π. Critical values are also rarely an issue, as they are often determined based on standard asymptotic results.

In our case, the problem comes from the empirical distribution of test statistics. Indeed, the quantities we compute are supposed to be distributed according to the Student or Gaussian distribution. However, because paths share similarities, correlation issues arise, as statistics are clearly not independent. This is a possible explanation to why distributions may be distorted.

D Additional figures

D.1 Baseline results

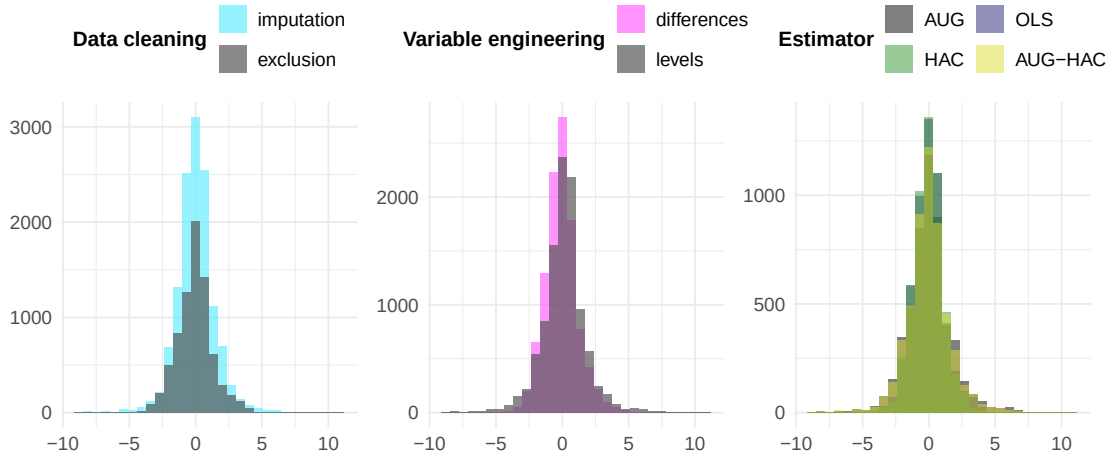

Figure 22: **Impact of mappings: robustness checks**. We report the distribution of t-statistics for two binary choices in mappings, plus the final estimator type. Results for regressions with fewer than 30 observations are discarded.

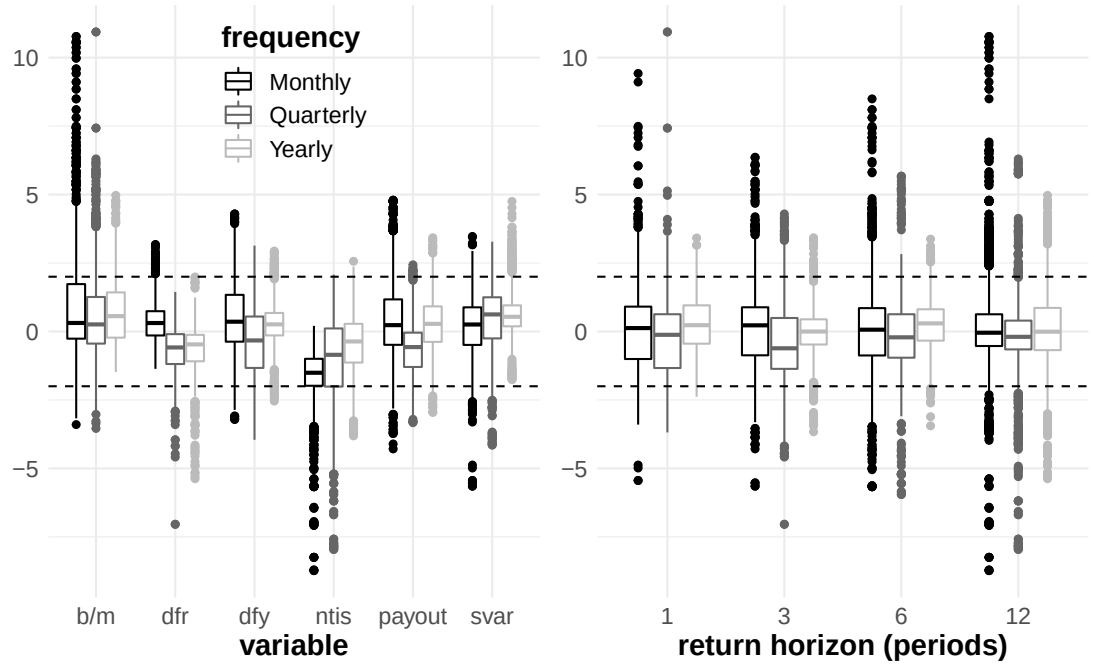

Figure 23: **Drivers of scalar output: modelling assumptions**. We report the distribution of t-statistics for two important modelling choices: the independent variable (left panel), and the return horizon of the dependent variable. Results for regressions with fewer than 30 observations are discarded.

D.2 Hacking intervals

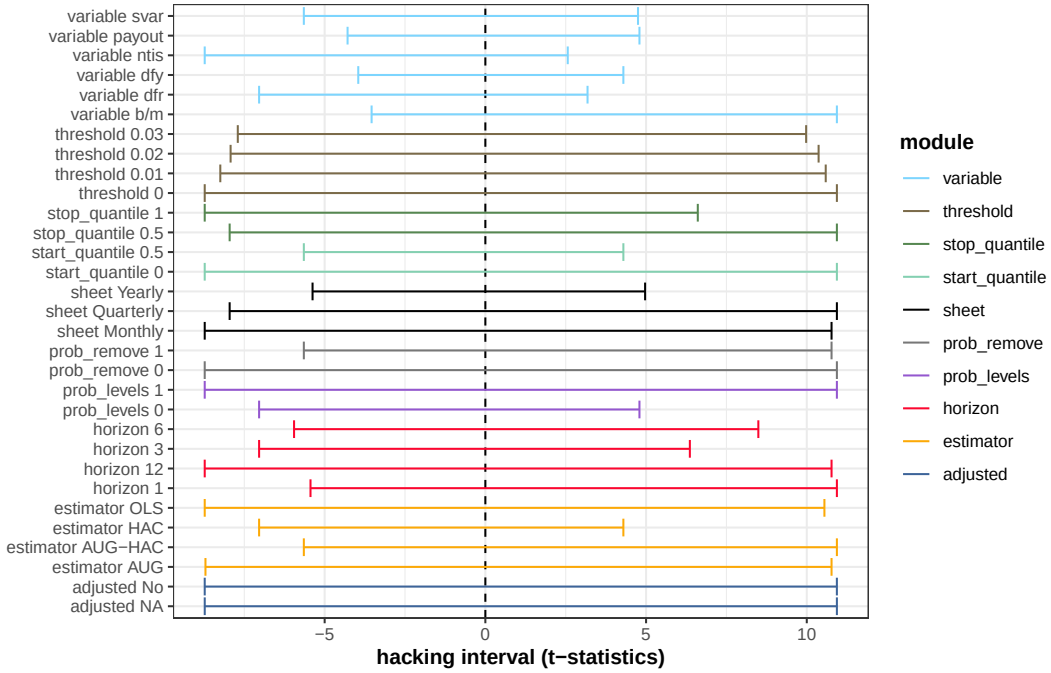

Figure 24: **Hacking intervals with one fixed mapping**. We show the intervals of t-statistics obtained when fixing one mapping. Each option of the mapping is tested and all combinations of all other mappings are spanned to generate the hacking intervals. The ten modules (i.e., mappings) are shown with colors.

D.3 Test assets and the magnitude of risk premia

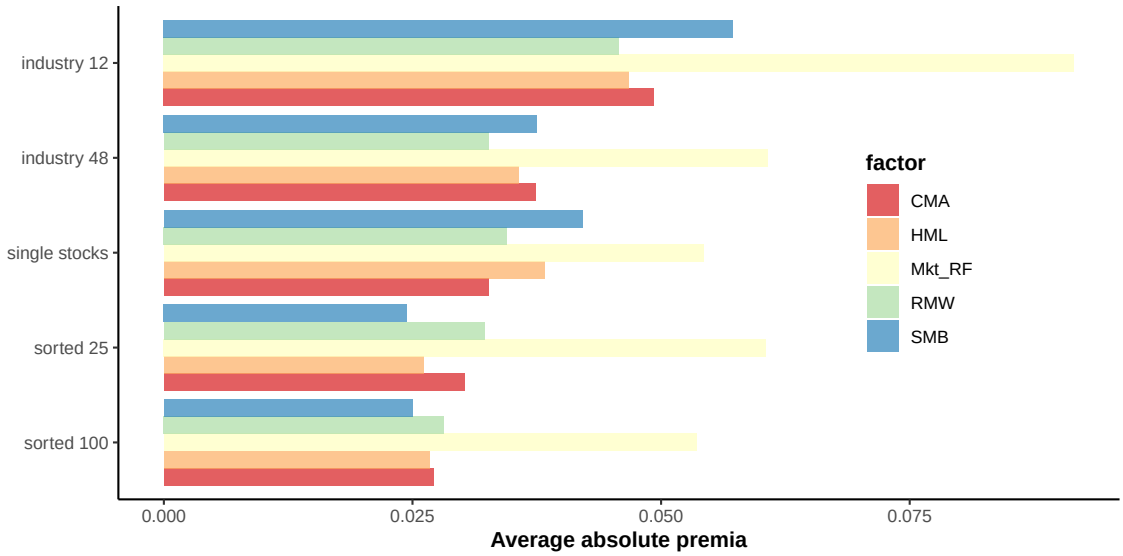

Figure 25: **Test asset and average abolute premia**. We plot the mean absolute premium (x-axis) of factors (shown with colors) for each group of test assets (y-axis).

References

Abadie, A. (2020). Statistical nonsignificance in empirical economics. *American Economic Review: Insights 2*(2), 193–208.

Akey, P., A. Robertson, and M. Simutin (2023). Noisy factors. *SSRN Working Paper 3930228*.

Amenc, N., F. Goltz, and B. Luyten (2020). Intangible capital and the value factor: Has your value definition just expired? *Journal of Portfolio Management 46*(7), 83–99.

Amihud, Y. and C. M. Hurvich (2004). Predictive regressions: A reduced-bias estimation method. *Journal of Financial and Quantitative Analysis 39*(4), 813–841.

Amrhein, V., S. Greenland, and B. McShane (2019). Scientists rise up against statistical significance. *Nature 567*, 305–307.

Andrews, D. W. and J. C. Monahan (1992). An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator. *Econometrica 60*(4), 953–966.

Andrews, I. and M. Kasy (2019). Identification of and correction for publication bias. *American Economic Review 109*(8), 2766–94.

Ang, A., J. Liu, and K. Schwarz (2020). Using stocks or portfolios in tests of factor models. *Journal of Financial and Quantitative Analysis 55*(3), 709–750.

Arjovsky, M., L. Bottou, I. Gulrajani, and D. Lopez-Paz (2019). Invariant risk minimization. *arXiv Preprint* (1907.02893).

Asness, C. and A. Frazzini (2013). The devil in HML's details. *Journal of Portfolio Management 39*(4), 49–68.

Asness, C. S., T. J. Moskowitz, and L. H. Pedersen (2013). Value and momentum everywhere. *Journal of Finance 68*(3), 929–985.

Avramov, D., S. Cheng, L. Metzker, and S. Voigt (2023). Integrating factor models. *Journal of Finance 78*(3), 1593–1646.

Azriel, D. and A. Schwartzman (2015). The empirical distribution of a large number of correlated normal variables. *Journal of the American Statistical Association 110*(511), 1217–1228.

Bailey, D. H., J. Borwein, M. Lopez de Prado, and Q. J. Zhu (2014). Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample performance. *Notices of the American Mathematical Society 61*(5), 458–471.

Bailey, D. H. and M. Lopez de Prado (2021). Finance is not excused: Why finance should not flout basic principles of statistics. *SSRN Working Paper 3895330*.

Bali, T. G., N. Cakici, and R. F. Whitelaw (2011). Maxing out: Stocks as lotteries and the cross-section of expected returns. *Journal of Financial Economics 99*(2), 427–446.

Bandi, F. M., B. Perron, A. Tamoni, and C. Tebaldi (2019). The scale of predictability. *Journal of Econometrics 208*(1), 120–140.

Barnett, A. G. and J. D. Wren (2019). Examination of cis in health and medical journals from 1976 to 2019: an observational study. *BMJ open 9*(11), e032506.

Barras, L., P. Gagliardini, and O. Scaillet (2022). Skill, scale, and value creation in the mutual fund industry. *Journal of Finance 77*(1), 601–638.

Barroso, P. and P. Santa-Clara (2015). Momentum has its moments. *Journal of Financial Economics 116*(1), 111–120.

Basit, T. (2003). Manual or electronic? The role of coding in qualitative data analysis. *Educational research 45*(2), 143–154.

Begg, C. B. and J. A. Berlin (1988). Publication bias: a problem in interpreting medical data. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 151*(3), 419–445.

Benjamin, D. J., J. O. Berger, M. Johannesson, B. A. Nosek, E.-J. Wagenmakers, R. Berk, K. A. Bollen, B. Brembs, L. Brown, C. Camerer, et al. (2018). Redefine statistical significance. *Nature Human Behaviour 2*(1), 6–10.

Bentkus, V., F. Götze, and A. Tikhomirov (1997). Berry-esseen bounds for statistics of weakly dependent samples. *Bernoulli 3*(3), 329–349.

Bergman, A., A. Chinco, S. M. Hartzmark, and A. B. Sussman (2020). Survey curious? Start-up guide and best practices for running surveys and experiments online. *SSRN Working Paper 3701330*.

Bessembinder, H., A. Burt, and C. M. Hrdlicka (2022). Factor returns and out-of-sample alphas: Factor construction matters. *SSRN Working Paper 4281769*.

Bhojraj, S. and B. Swaminathan (2006). Macromomentum: returns predictability in international equity indices. *Journal of Business 79*(1), 429–451.

Blanco-Perez, C. and A. Brodeur (2020). Publication bias and editorial statement on negative findings. *Economic Journal 130*(629), 1226–1247.

Boissel, C. and A. Matray (2022). Dividend taxes and the allocation of capital. *American Economic Review 112*(9), Retracted.

Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze 8*, 3–62.

Boriah, S., V. Chandola, and V. Kumar (2008). Similarity measures for categorical data: A comparative evaluation. In *Proceedings of the 2008 SIAM International Conference on Data Mining*, pp. 243–254. SIAM.

Boudoukh, J., R. Israel, and M. Richardson (2022). Biases in long-horizon predictive regressions. *Journal of Financial Economics 145*(3), 937–969.

Boudoukh, J., M. Richardson, and R. F. Whitelaw (2008). The myth of long-horizon predictability. *Review of Financial Studies 21*(4), 1577–1605.

Boylan, J. E. (2016). Reproducibility. *IMA Journal of Management Mathematics 27*(2), 107–108.

Breznau, N., E. M. Rinke, A. Wuttke, H. H. Nguyen, M. Adem, J. Adriaans, A. Alvarez-Benjumea, H. K. Andersen, D. Auer, F. Azevedo, et al. (2022). Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *Proceedings of the National Academy of Sciences 119*(44), e2203150119.

Brodeur, A., N. Cook, and A. Heyes (2020). Methods matter: P-hacking and publication bias in causal analysis in economics. *American Economic Review 110*(11), 3634–60.

Brodeur, A., N. Cook, and A. Heyes (2022). We need to talk about mechanical turk: What 22,989 hypothesis tests tell us about publication bias and p-hacking in online experiments. *SSRN Working Paper 4188289*.

Brodeur, A., M. Lé, M. Sangnier, and Y. Zylberberg (2016). Star wars: The empirics strike back. *American Economic Journal: Applied Economics 8*(1), 1–32.

Buckland, S. T., K. P. Burnham, and N. H. Augustin (1997). Model selection: an integral part of inference. *Biometrics 53*(2), 603–618.

Bühlmann, P. (2020). Invariance, causality and robustness. *Statistical Science 35*(3), 404–426.

Burnham, K. and D. Anderson (2004a). *Model selection and multimodel inference*. Springer-Verlag. 2nd Edution.

Burnham, K. P. and D. R. Anderson (2004b). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research 33*(2), 261–304.

Camerer, C. F., A. Dreber, E. Forsell, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, J. Almenberg, A. Altmejd, and T. Chan (2016). Evaluating replicability of laboratory experiments in economics. *Science 351*(6280), 1433–1436.

Carver, R. (1978). The case against statistical significance testing. *Harvard Educational Review 48*(3), 378–399.

Chan, K., A. Hameed, and W. Tong (2000). Profitability of momentum strategies in the international equity markets. *Journal of Financial and Quantitative Analysis 35*(2), 153–172.

Chang, X. S., H. Gao, and W. Li (2023). Discontinuous distribution of test statistics around significance thresholds in empirical accounting studies. *SSRN Working Paper 3762342*.

Chen, A. Y. (2021). Most claimed statistical findings in cross-sectional return predictability are likely true. *SSRN Working Paper 3912915*.

Chen, A. Y., A. Lopez-Lira, and T. Zimmermann (2023). Peer-reviewed theory does not help predict the cross-section of stock returns. *arXiv Preprint* (2212.10317).

Chen, A. Y. and M. Velikov (2023). Zeroing in on the expected returns of anomalies. *Journal of Financial and Quantitative Analysis 58*(3), 968–1004.

Chen, A. Y. and T. Zimmermann (2022a). Open source cross-sectional asset pricing. *Critical Finance Review 27*(2), 207–264.

Chen, A. Y. and T. Zimmermann (2022b). Publication bias in asset pricing research. *arXiv Preprint* (2209.13623).

Chen, L. and W. B. Wu (2018). Concentration inequalities for empirical processes of linear time series. *J. Mach. Learn. Res. 18*, 1–46.

Chen, M., C. Gao, and Z. Ren (2016). A general decision theory for huber's ϵ-contamination model. *Electronic Journal of Statistics 10*, 3752–3774.

Chinco, A., A. Neuhierl, and M. Weber (2021). Estimating the anomaly base rate. *Journal of Financial Economics 140*(1), 101–126.

Chopra, F., I. Haaland, C. Roth, and A. Stegmann (2022). The null result penalty. *SSRN Working Paper 4144648*.

Chordia, T., A. Goyal, and A. Saretto (2020). Anomalies and false rejections. *Review of Financial Studies 33*(5), 2134–2179.

Christensen, G. and E. Miguel (2018). Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature 56*(3), 920–80.

Coker, B., C. Rudin, and G. King (2021). A theory of statistical inference for ensuring the robustness of scientific results. *Management Science 67*(10), 6174–6197.

Collaboration, O. S. (2015). Estimating the reproducibility of psychological science. *Science 349*(6251), aac4716.

Colliard, J.-E., C. Hurlin, and C. Pérignon (2022). The economics of computational reproducibility. *SSRN Working Paper 3418896*.

Cortina, J. M., T. Koehler, and L. Craig (2022). Are we building our science on quicksand? Current reproducibility practices in management. In *Academy of Management Proceedings*, Volume 2022, pp. 15461.

Dai, R., L. Donohue, Q. F. S. Drechsler, and W. Jiang (2023). Dissemination, publication, and impact of finance research: When novelty meets conventionality. *Review of Finance 27*(1), 79–141.

Daniel, K. and T. J. Moskowitz (2016). Momentum crashes. *Journal of Financial Economics 122*(2), 221–247.

De Long, J. B. and K. Lang (1992). Are all economic hypotheses false? *Journal of Political Economy 100*(6), 1257–1272.

De Prado, M. L. (2018). The 10 reasons most machine learning funds fail. *Journal of Portfolio Management 44*(6), 120–133.

de Prado, M. L. (2023). The hierarchy of empirical evidence in finance. *Journal of Portfolio Management 49*(9), 10–29.

Demmel, J. (1992). The componentwise distance to the nearest singular matrix. *SIAM Journal on Matrix Analysis and Applications 13*(1), 10–19.

Diananda, P. (1955). The central limit theorem for m-dependent variables. In *Mathematical Proceedings of the Cambridge Philosophical Society*, Volume 51, pp. 92–95.

Dichtl, H., W. Drobetz, A. Neuhierl, and V.-S. Wendt (2021). Data snooping in equity premium prediction. *International Journal of Forecasting 37*(1), 72–94.

Dickersin, K., S. Chan, T. Chalmersx, H. Sacks, and H. Smith Jr (1987). Publication bias and clinical trials. *Controlled Clinical Trials 8*(4), 343–353.

Doucouliagos, C. and T. D. Stanley (2013). Are all economic facts greatly exaggerated? Theory competition and selectivity. *Journal of Economic Surveys 27*(2), 316–339.

Doucouliagos, H. and T. D. Stanley (2009). Publication selection bias in minimum-wage research? A meta-regression analysis. *British Journal of Industrial Relations 47*(2), 406–428.

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society: Series B (Methodological) 57*(1), 45–70.

Duvendack, M., R. Palmer-Jones, and W. R. Reed (2017). What is meant by" replication" and why does it encounter resistance in economics? *American Economic Review 107*(5), 46–51.

Echenique, F. and K. He (2023). Screening p-hackers: Dissemination noise as bait. *arXiv Preprint* (2103.09164).

Elliott, G., N. Kudrin, and K. Wuthrich (2022). Detecting p-hacking. *Econometrica 90*(2), 887–906.

Engelberg, J., R. D. McLean, J. Pontiff, and M. C. Ringgenberg (2023). Do cross-sectional predictors contain systematic information? *Journal of Financial and Quantitative Analysis 58*(3), 1172–1201.

Fabozzi, F. J. and M. L. de Prado (2018). Being honest in backtest reporting: A template for disclosing multiple tests. *Journal of Portfolio Management 45*(1), 141–147.

Fama, E. F. and K. R. French (1992). The cross-section of expected stock returns. *Journal of Finance 47*(2), 427–465.

Fama, E. F. and K. R. French (2015). A five-factor asset pricing model. *Journal of Financial Economics 116*(1), 1–22.

Fama, E. F. and J. D. MacBeth (1973). Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy 81*(3), 607–636.

Fan, J. and X. Han (2017). Estimation of the false discovery proportion with unknown dependence. *Journal of the Royal Statistical Society. Series B, Statistical methodology 79*(4), 1143.

Fanelli, D. (2018). Opinion: Is science really facing a reproducibility crisis, and do we need it to? *Proceedings of the National Academy of Sciences 115*(11), 2628–2631.

Fanelli, D., R. Costas, and J. P. Ioannidis (2017). Meta-assessment of bias in science. *Proceedings of the National Academy of Sciences 114*(14), 3714–3719.

Farcomeni, A. (2008). A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Statistical methods in medical research 17*(4), 347–388.

Farmer, L., L. Schmidt, and A. Timmermann (2023). Pockets of predictability. *Journal of Finance 78*(3), 1279–1341.

Feng, G., S. Giglio, and D. Xiu (2020). Taming the factor zoo: A test of new factors. *Journal of Finance 75*(3), 1327–1370.

Fernandez, C., E. Ley, and M. F. Steel (2001). Benchmark priors for bayesian model averaging. *Journal of Econometrics 100*(2), 381–427.

Fisher, A., C. Rudin, and F. Dominici (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res. 20*(177), 1–81.

Frankel, A. and M. Kasy (2022). Which findings should be published? *American Economic Journal: Microeconomics 14*(1), 1–38.

Gagliardini, P., E. Ossola, and O. Scaillet (2016). Time-varying risk premium in large cross-sectional equity data sets. *Econometrica 84*(3), 985–1046.

Gelman, A. and E. Loken (2014). The statistical crisis in science. *American Scientist 102*, 460–465.

Giglio, S., Y. Liao, and D. Xiu (2021). Thousands of alpha tests. *Review of Financial Studies 34*(7), 3456–3496.

Golub, G. H. and J. H. Wilkinson (1966). Note on the iterative refinement of least squares solution. *Numerische Mathematik 9*(2), 139–148.

Gong, Q., M. Liu, and Q. Liu (2015). Momentum is really short-term momentum. *Journal of Banking & Finance 50*, 169–182.

Gould, E., H. S. Fraser, T. H. Parker, S. Nakagawa, S. C. Griffith, P. A. Vesk, F. Fidler, R. N. Abbey-Lee, J. K. Abbott, L. A. Aguirre, et al. (2023). Same data, different analysts: variation in effect sizes due to analytical decisions in ecology and evolutionary biology.

Goyal, A., I. Welch, and A. Zafirov (2023). A comprehensive 2022 look at the empirical performance of equity premium prediction. *SSRN Working Paper 3929119*.

Granger, C. W. and H. F. Uhlig (1990). Reasonable extreme-bounds analysis. *Journal of Econometrics 44*(1-2), 159–170.

Grcar, J. F. (2003). Optimal sensitivity analysis of linear least squares. *Lawrence Berkeley National Laboratory, Report LBNL-52434 99*.

Greene, W. H. (2018). *Econometric analysis - Eighth Edition*. Pearson Education India.

Griffin, J. M., X. Ji, and J. S. Martin (2003). Momentum investing and business cycle risk: Evidence from pole to pole. *Journal of Finance 58*(6), 2515–2547.

Gu, S., B. Kelly, and D. Xiu (2020). Empirical asset pricing via machine learning. *Review of Financial Studies 33*(5), 2223–2273.

Hansen, B. E. (2007). Least squares model averaging. *Econometrica 75*(4), 1175–1189.

Hansen, P. R., A. Lunde, and J. M. Nason (2011). The model confidence set. *Econometrica 79*(2), 453–497.

Harvey, C. R. (2017). Presidential address: The scientific outlook in financial economics. *Journal of Finance 72*(4), 1399–1440.

Harvey, C. R. (2021). Be skeptical of asset management research. *Available at SSRN Working Paper 3906277*.

Harvey, C. R. and Y. Liu (2020). False (and missed) discoveries in financial economics. *Journal of Finance 75*(5), 2503–2553.

Harvey, C. R. and Y. Liu (2021a). Lucky factors. *Journal of Financial Economics 141*(2), 413–435.

Harvey, C. R. and Y. Liu (2021b). Uncovering the iceberg from its tip: A model of publication bias and p-hacking. *SSRN Working Paper 3865813*.

Harvey, C. R., Y. Liu, and A. Saretto (2020). An evaluation of alternative multiple testing methods for finance applications. *Review of Asset Pricing Studies 10*(2), 199–248.

Harvey, C. R., Y. Liu, and H. Zhu (2016). ... and the cross-section of expected returns. *Review of Financial Studies 29*(1), 5–68.

Head, M. L., L. Holman, R. Lanfear, A. T. Kahn, and M. D. Jennions (2015). The extent and consequences of p-hacking in science. *PLoS Biology 13*(3), e1002106.

Hjalmarsson, E. (2011). New methods for inference in long-horizon regressions. *Journal of Financial and Quantitative Analysis 46*(3), 815–839.

Hoeffding, W. and H. Robbins (1948). The central limit theorem for dependent random variables. *Duke Mathematical Journal 15*(3), 773–780.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999). Bayesian model averaging: A tutorial. *Statistical Science 14*(4), 382–417.

Hofman, J. M., D. G. Goldstein, S. Sen, F. Poursabzi-Sangdeh, J. Allen, L. L. Dong, B. Fried, H. Gaur, A. Hoq, E. Mbazor, et al. (2021). Expanding the scope of reproducibility research through data analysis replications. *Organizational Behavior and Human Decision Processes 164*, 192–202.

Hollenbeck, J. R. and P. M. Wright (2017). Harking, sharking, and tharking: Making the case for post hoc analysis of scientific data.

Hollstein, F., M. Prokopczuk, and V. Voigts (2022). How robust are empirical factor models to the choice of breakpoints? *SSRN Working Paper 3924821*.

Hou, K., C. Xue, and L. Zhang (2020). Replicating anomalies. *Review of Financial Studies 33*(5), 2019–2133.

Huber, C., A. Dreber, J. Huber, M. Johannesson, M. Kirchler, U. Weitzel, M. Abellán, X. Adayeva, F. C. Ay, K. Barron, et al. (2023). Competition and moral behavior: A meta-analysis of forty-five crowd-sourced experimental designs. *Proceedings of the National Academy of Sciences 120*(23), e2215572120.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics 35*(1), 73–101.

Huntington-Klein, N., A. Arenas, E. Beam, M. Bertoni, J. R. Bloem, P. Burli, N. Chen, P. Grieco, G. Ekpe, and T. Pugatch (2021). The influence of hidden researcher decisions in applied microeconomics. *Economic Inquiry 59*, 944–960.

Imbens, G. W. (2021). Statistical significance, *p*-values, and the reporting of uncertainty. *Journal of Economic Perspectives 35*(3), 157–74.

Ioannidis, J., T. Stanley, and H. Doucouliagos (2017). The power of bias in economics research. *Economic Journal 127*(605), F236–F265.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine 2*(8), e124.

Jegadeesh, N. and S. Titman (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *Journal of Finance 48*(1), 65–91.

Jensen, T. I., B. T. Kelly, and L. H. Pedersen (2021). Is there a replication crisis in finance? *Journal of Finance Forthcoming*.

Jirak, M. (2016). Berry–esseen theorems under weak dependence. *Annals of Probability 44*(3), 2024–2063.

Jirak, M. (2023). A berry-esseen bound with (almost) sharp dependence conditions. *Bernoulli 29*(2), 1219–1245.

Jørgensen, T. H. (2023). Sensitivity to calibrated parameters. *Review of Economics and Statistics 105*(2), 474–481.

Kapoor, S. and A. Narayanan (2022). Leakage and the reproducibility crisis in ml-based science. *arXiv Preprint 2207.07048*.

Kasy, M. (2021). Of forking paths and tied hands: Selective publication of findings, and what economists should do about it. *Journal of Economic Perspectives 35*(3), 175–92.

Kerr, N. L. (1998). Harking: Hypothesizing after the results are known. *Personality and social psychology review 2*(3), 196–217.

Kjelsrud, A., A. Kotsadam, and O. Rogeberg (2023). Cooperative property rights and development: Evidence from land reform in El Salvador: A Comment. *Journal of Political Economy Forthcoming*(1).

Kontorovich, A. and R. Weiss (2014). Uniform Chernoff and Dvoretzky-Kiefer-Wolfowitz-type inequalities for markov chains and related processes. *Journal of Applied Probability 51*(4), 1100–1113.

Leamer, E. and H. Leonard (1983). Reporting the fragility of regression estimates. *Review of Economics and Statistics 65*(2), 306–317.

Leamer, E. E. (1985). Sensitivity analyses would help. *American Economic Review 75*(3), 308–313.

Leek, J. T. and L. R. Jager (2017). Is most published research really false? *Annual Review of Statistics and Its Application 4*, 109–122.

Lo, A. W. and A. C. MacKinlay (1990). Data-snooping biases in tests of financial asset pricing models. *Review of Financial Studies 3*(3), 431–467.

Loh, P.-L. and X. L. Tan (2018). High-dimensional robust precision matrix estimation: Cellwise corruption under ϵ-contamination. *Electronic Journal of Statistics 12*(1), 1429–1467.

Mariette, J., O. Blanchard, O. Berné, and T. B. Ari (2021). An open-source tool to assess the carbon footprint of research. *arXiv Preprint* (2101.10124).

McCloskey, A. and P. Michaillat (2023). Critical values robust to p-hacking. *arXiv Preprint* (2005.04141).

McShane, B. B., D. Gal, A. Gelman, C. Robert, and J. L. Tackett (2019). Abandon statistical significance. *American Statistician 73*, 235–245.

Menkveld, A., A. Dreber, F. Holzmeister, M. Johannesson, J. Huber, M. Kirchler, S. Neususs, M. Razen, and U. Weitzel (2023). Non-standard errors. *Journal of Finance Forthcoming*.

Milkman, K. et al. (2021). Megastudies improve the impact of applied behavioral science. *Nature 600*, 478–483.

Mitton, T. (2022). Methodological variation in empirical corporate finance. *Review of Financial Studies 35*(2), 527–575.

Monahan, T. and J. A. Fisher (2010). Benefits of 'observer effects': lessons from the field. *Qualitative research 10*(3), 357–376.

Moral-Benito, E. (2015). Model averaging in economics: An overview. *Journal of Economic Surveys 29*(1), 46–75.

Morey, M. R. and S. Yadav (2018). Documentation of the file drawer problem in academic finance journals. *Journal of Investing 27*(1), 143–147.

Moskowitz, T. J. and M. Grinblatt (1999). Do industries explain momentum? *Journal of Finance 54*(4), 1249–1290.

Mueller-Langer, F., B. Fecher, D. Harhoff, and G. G. Wagner (2019). Replication studies in economics—how many and which papers are chosen for replication, and why? *Research Policy 48*(1), 62–83.

Muralidharan, K., M. Romero, and K. Wüthrich (2022). Factorial designs, model selection, and (incorrect) inference in randomized experiments. *SSRN Working Paper 3551804*.

Newey, W. and K. West (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica 55*(3), 703–708.

Novy-Marx, R. (2012). Is momentum really momentum? *Journal of Financial Economics 103*(3), 429–453.

Olson, C. M., D. Rennie, D. Cook, K. Dickersin, A. Flanagin, J. W. Hogan, Q. Zhu, J. Reiling, and B. Pace (2002). Publication bias in editorial decision making. *Journal of the American Medical Association 287*(21), 2825–2828.

Orey, S. (1958). A central limit theorem for m-dependent random variables. *Duke Mathematical Journal 25*(4), 543–546.

Pérignon, C., O. Akmansoy, C. Hurlin, A. Dreber, F. Holzmeister, J. Huber, M. Johannesson, M. Kirchler, A. J. Menkveld, M. Razen, et al. (2023). Computational reproducibility in finance: Evidence from 1,000 tests. *SSRN Working Paper 4064172*.

Peters, J., P. Bühlmann, and N. Meinshausen (2016). Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 78*(5), 947–1012.

Peters, J., N. Fiala, and F. Neubauer (2023). Do economists replicate? *Journal of Economic Behavior and Organization 212*, 219–232.

Pfister, N., P. Bühlmann, and J. Peters (2019). Invariant causal prediction for sequential data. *Journal of the American Statistical Association 114*(527), 1264–1276.

Raftery, A. E., D. Madigan, and J. A. Hoeting (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association 92*(437), 179–191.

Rampini, A. A., S. Viswanathan, and G. Vuillemey (2021). Risk management in financial institutions. *Journal of Finance 75*(2), Retracted.

Romano, J. P. and M. Wolf (2005). Stepwise multiple testing as formalized data snooping. *Econometrica 73*(4), 1237–1282.

Romano, J. P. and M. Wolf (2010). Balanced control of generalized error rates. *Annals of Statistics 38*(1), 598–633.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological bulletin 86*(3), 638–641.

Rouwenhorst, K. G. (1998). International momentum strategies. *Journal of Finance 53*(1), 267–284.

Rytchkov, O. and X. Zhong (2020). Information aggregation and p-hacking. *Management Science 66*(4), 1605–1626.

Serra-Garcia, M. and U. Gneezy (2021). Nonreplicable publications are cited more than replicable ones. *Science Advances 7*(21), eabd1705.

Shepperd, M. and L. Yousefi (2023). An analysis of retracted papers in computer science. *Plos one 18*(5), e0285383.

Simonsohn, U., L. D. Nelson, and J. P. Simmons (2014a). P-curve: a key to the file-drawer. *Journal of Experimental Psychology: General 143*(2), 534.

Simonsohn, U., L. D. Nelson, and J. P. Simmons (2014b). p-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science 9*(6), 666–681.

Simonsohn, U., J. P. Simmons, and L. D. Nelson (2020). Specification curve analysis. *Nature Human Behaviour 4*(11), 1208–1214.

Smith, S. C. and A. Timmermann (2022). Have risk premia vanished? *Journal of Financial Economics 145*(2), 553–576.

Soebhag, A., B. van Vliet, and P. Verwijmeren (2023). Non-standard errors in asset pricing: Mind your sorts. *SSRN Working Paper 4136672*.

Stambaugh, R. F. (1999). Predictive regressions. *Journal of Financial Economics 54*(3), 375–421.

Stanley, T. D. (2005). Beyond publication bias. *Journal of Economic Surveys 19*(3), 309–345.

Steel, M. F. (2020). Model averaging and its use in economics. *Journal of Economic Literature 58*(3), 644–719.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association 54*(285), 30–34.

Stroebe, W. (2019). What can we learn from many labs replications? *Basic and Applied Social Psychology 41*(2), 91–103.

Van Aert, R. C., J. M. Wicherts, and M. A. Van Assen (2019). Publication bias examined in meta-analyses from psychology and medicine: A meta-meta-analysis. *PloS One 14*(4), e0215052.

van Zwet, E. W. and E. A. Cator (2021). The significance filter, the winner's curse and the need to shrink. *Statistica Neerlandica 75*(4), 437–452.

Vilhuber, L. (2020). Reproducibility and replicability in economics. *Harvard Data Science Review 2*(4).

Viviano, D., K. Wuthrich, and P. Niehaus (2022). (When) should you adjust inferences for multiple hypothesis testing? *arXiv Preprint* (2104.13367).

Vu, P. (2022). Can the replication rate tell us about publication bias? *arxiv Preprint* (2206.15023).

Walter, D., R. Weber, and P. Weiss (2023). Non-standard errors in portfolio sorts. *SSRN Working Paper 4164117*.

Wang, J., Q. Zhao, T. Hastie, and A. B. Owen (2017). Confounder adjustment in multiple hypothesis testing. *Annals of Statistics 45*(5), 1863.

Wasserstein, R. L., A. L. Schirm, and N. A. Lazar (2019). Moving to a world beyond "p< 0.05". *American Statistician 73*(sup1), 1–19.

Welch, I. and A. Goyal (2008). A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies 21*(4), 1455–1508.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 817–838.

White, H. (1996). *Estimation, inference and specification analysis*. Number 22. Cambridge University Press.

White, H. (2000). A reality check for data snooping. *Econometrica 68*(5), 1097–1126.

Wickham, H. (2014). Tidy data. *Journal of Statistical Software 59*(1), 1–23.

Wilson, D. J. (2019). The harmonic mean p-value for combining dependent tests. *Proceedings of the National Academy of Sciences 116*(4), 1195–1200.

Woolston, C. (2015). Psychology journal bans p values. *Nature 519*(7541), 9–9.

Yan, X. S. and L. Zheng (2017). Fundamental analysis and the cross-section of stock returns: A data-mining approach. *Review of Financial Studies 30*(4), 1382–1423.

Zhang, X. (2015). Consistency of model averaging estimators. *Economics Letters 130*, 120–123.

Zhang, X. and C.-A. Liu (2019). Inference after model averaging in linear regression models. *Econometric Theory 35*(4), 816–841.

Zhu, R., X. Zhang, A. T. Wan, and G. Zou (2023). Kernel averaging estimators. *Journal of Business & Economic Statistics 41*(1), 157–169.

Ziliak, S. and D. N. McCloskey (2008). *The cult of statistical significance: How the standard error costs us jobs, justice, and lives*. University of Michigan Press.