Persistent Anomalies and Nonstandard Errors

Guillaume Coqueret* Christophe Pérignon†

October 19, 2025

Abstract

We develop a framework for rigorous inference when assessing asset pricing anomalies and accounting for multiple methodological choices. We demonstrate that running multiple paths on the same dataset results in high correlation across outcomes, biasing inference. Alternatively, path-specific resampling reduces outcome correlations and tightens the confidence interval of the average return. Accounting for across and within-path variability allows us to decompose the variance of the average return into a standard error, a nonstandard error, and a correlation term. Empirically, we identify 29 persistent anomalies with statistically significant average returns and show that, for most anomalies, nonstandard errors dominate standard errors.

Keywords: Asset pricing anomalies, p-hacking, multi-path inference, resampling, research replicability, nonstandard errors

JEL: C12, C18, C51, G12

*EMLYON Business School, 144, avenue Jean Jaures, 69007 Lyon, France. ✉ coqueret@em-lyon.com.
†HEC Paris, 1 Rue de la Libération, 78350 Jouy-en-Josas, France. ✉ perignong@hec.fr.

1 Introduction

The finance literature has recently shown increasing interest in multi-design studies (see Table 1), building on emerging practices in other scientific disciplines.[1] These empirical studies consider numerous variations of the baseline methodology, often referred to as forking paths. They can be either conducted by multiple independent research teams (multi-analyst studies, e.g., Menkveld et al. (2024)) or by a single research team considering various potential modeling decisions (multi-path studies, e.g., Soebhag et al. (2024)).

Multi-design studies provide two distinct advantages. First, by estimating a distribution of effects rather than a single point estimate, they provide a more comprehensive characterization of the analyzed phenomenon and serve as a potential remedy for p-hacking in empirical research (Chen, 2021).[2] Second, they quantify the uncertainty arising from ad hoc methodological choices made by researchers, which Menkveld et al. (2024) coined as *nonstandard errors* (NSE).

In the context of asset pricing, multi-design studies open up valuable opportunities. Indeed, the debate surrounding the robustness of empirical findings and the uncertainty stemming from methodological decisions is particularly intense regarding the so-called asset-pricing *anomalies* (Fama and French, 1996; Hou et al., 2015; McLean and Pontiff, 2016). Given the proliferation of these return regularities—statistically significant, persistent, and unexplained by standard risk-based models—and their practical importance in the asset management industry, there is a pressing need for robust approaches to navigate the "factor zoo" (Harvey et al., 2016; Feng et al., 2020; Bryzgalova et al., 2023).

Despite growing interest in multi-design studies, rigorous approaches for handling the large number of resulting estimates remain underdeveloped. In this paper, we propose a framework that enables two methodological contributions to the literature. First, we demonstrate how to formally test whether the average return of anomaly-based portfolios differs from zero. We derive confidence intervals for identifying *persistent anomalies*—those that are robust to multiple methodological variations and to data resampling. Second, we decompose the total variance of the average return into two distinct sources: the standard error (SE) capturing the variability arising from sampling and the NSE capturing the variability attributable to differences across paths. In contrast to the existing literature, (the square of) our estimates for the SE and NSE of the effect sums exactly to the total variance of the effect.

The various steps of our analysis are the following. We start by showing that overlapping methodological paths applied to the same dataset inherently produce highly correlated estimates, with correlation coefficients exhibiting a skewed distribution. We then show how this strong correlation structure across outcomes distorts inference, resulting in wide intervals that hinder our ability to draw definitive conclusions about the sign and

[1] Key references include Gelman and Loken (2014) in statistics, Silberzahn et al. (2018) in psychology, Botvinik-Nezer et al. (2020) in machine learning, Huntington-Klein et al. (2021), Breznau et al. (2022, 2024) and Huntington-Klein et al. (2025) in economics, Gould et al. (2023) in biology, and Huber et al. (2023) in behavioral sciences.

[2] P-hacking corresponds to relentless analysis of data with an intent to obtain a statistically significant result, usually to support the researcher's hypothesis (Elliott et al., 2022; Brodeur et al., 2016).

intensity of the effect. This shares similarities with the autocorrelation corrections (HAC) that have been proposed in the time-series literature (Newey and West (1987)). In our setting, however, the correlation arises not from temporal ordering but from the cross-sectional dependence across paths. Consequently, we argue that the variance calibration should be adjusted in a manner analogous to HAC estimators.

At first glance, the idea that highly correlated estimates could be problematic may seem counterintuitive. After all, if a finding is truly robust, would not different paths naturally become correlated as they are capturing the same effect? The benefits of low correlations have long been documented in multiple testing (Benjamini and Hochberg (1995)) or in model combination, e.g., bagging and ensembles (see Breiman (1996) or Zhou (2025)). Recently, ensembles have been found to be a promising avenue in asset pricing, especially when they aggregate complementary models that learn different perspectives from the data (Kelly and Malamud (2025)). In our framework, each path functions like a separate model that independently learns from the data. Ideally, all paths should reach the same conclusion—but for different reasons. When such agreement emerges from diverse perspectives, it signals robustness. Conversely, if the paths agree for the same reason (i.e., they are highly correlated), they contribute little additional information, effectively acting as a single path.

To mitigate the correlation effects, we demonstrate that randomly sampling data before running each path significantly reduces outcome correlations and symmetrizes their distribution around zero, leading to tighter confidence intervals. To see why this happens, we recall that the width of the confidence intervals around the mean, $\hat{\mu}_b$, increases with the variance of the estimated mean, $\sigma^2_{\hat{\mu}_b}$. We show that this variance is the product of two important terms. The first one is the average of correlations across all outcomes and the second one is the variance of the effect under study, σ^2_b. We rely on the law of total variance to decompose the variance of the effects into two components: $\sigma^2_b = \text{SE}^2 + \text{NSE}^2$. This novel decomposition allows us to derive a canonical expression for the variance of the mean effect: $\sigma^2_{\hat{\mu}_b} = (\text{SE}^2 + \text{NSE}^2) \sum_{p,q} \frac{\rho_{p,q}}{P^2}$, where $\rho_{p,q}$ denotes the correlation between the outcomes of paths p and q, and P is the total number of paths.

In an empirical analysis of 33 asset pricing anomalies, we consider seven critical methodological choices (e.g., sample period, holding period, long-short quantiles) for a total of 576 methodological paths. For each anomaly and each path, we estimate the average return of a long-short portfolio sorted on the corresponding firm characteristic. To compute the correlations among outcomes, we contrast two resampling strategies: (1) all paths are run on the same new samples (*common resampling*); and (2) new samples are drawn separately for each individual path (*specific resampling*). Note that this second approach is similar in spirit to the sampling of trees in random forests (Ho (1998); Breiman (2001)).

We show that the resulting average correlation across paths with common resampling is around 30%, whereas it is below 0.25% with the specific resampling we recommend. From the canonical decomposition of the variance, we directly see that the advantage of resorting to specific resampling is to shrink the variance more than 100 times. Consequently, the confidence interval for the mean effect, which is proportional to the standard deviation, is reduced approximately by a factor of ten. Applying our strategy to a large

sample of asset pricing anomalies reveals 29 that can be classified as persistent, with the strongest effects linked to trend-following and momentum strategies.

Our empirical study also provides some insights into the respective contributions of the SE and NSE to the variance of the (mean) effect. A robust finding of our study is that the NSE component dwarfs the SE component for most anomalies. This means that the bulk of the uncertainty concerning the performance of anomaly-based portfolios is due to methodological variation. In a set of extensions, we demonstrate that our framework (1) can incorporate non-uniform weighting schemes across paths to reflect preferences or theoretical guidance, (2) can be applied with alternative sampling methods, and (3) enables richer robustness checks in empirical finance applications.

Our paper adds to the literature on the validity, robustness, and credibility of empirical results in finance (Harvey, 2017). A first stream of the literature has focused on the *internal validity* of empirical findings. To make causal claims, finance researchers have exploited natural experiments and other sharp identification strategies, e.g., difference-in-differences, instrumental variables, and regression discontinuity design (Roberts and Whited, 2013; Heath et al., 2023). They have also accounted for multiple hypothesis testing and false discoveries to ensure that statistically significant results are not merely due to chance (Barras et al., 2010; Harvey and Liu, 2020; Chordia et al., 2020).

A second stream of the literature, more closely related to the present paper, has focused on the *external validity* of empirical findings. In his AFA Presidential Address, Harvey (2017) emphasizes the value of reanalysis studies in finance, arguing that they strengthen the field's scientific foundations and help build credibility (also see Nagel (2019)). Using the original code and data provided by the authors, Pérignon et al. (2024) independently verify the empirical results of a sample of finance research papers and report a reproducibility success rate of 52%. Over the past decade, several replication studies have challenged the robustness of some classic empirical results in corporate finance (Mitton, 2022; Cohn et al., 2023) and in asset pricing for equity returns (McLean and Pontiff, 2016; Harvey et al., 2016; Hou et al., 2020) and bond returns (Dickerson et al., 2023; Dick-Nielsen et al., 2023). In contrast, Jensen et al. (2023) and Chen and Zimmermann (2022b) successfully replicated the findings of a large number of asset pricing anomaly papers. To promote comparability and replicability, Hellum et al. (2025) design a collaborative and competitive process that allows the future performance of many asset pricing models to be evaluated on equal footing using a common, secret dataset.

The multi-design approach serves as a valuable complement to traditional reanalyses. Instead of relying on subsequent studies, often published years later to reassess the validity of existing results by, for instance, altering the sample period, outlier management, or the estimation method, multi-design studies aim to internalize methodological uncertainty by systematically spanning a range of protocol choices. While multi-design analyses focus on the role of methodological variability, we show in this paper that sampling variability also matters. Indeed, one must keep in mind that any given sample is only a single realization of the data-generating process. When running multiple protocols on a single sample, the analyst tends to overlook the importance of sampling noise. In contrast, we demonstrate that allowing for shifts in the baseline dataset improves inference in multi-design settings.

2 Multi-design methodologies

2.1 Current methodologies

We list in Table 1 recent studies that leverage multi-design analyses in the field of finance. These studies span topics such as portfolio strategy performance, market microstructure, and corporate finance, highlighting the broad relevance of multi-design approaches throughout the discipline. The table provides the number of methodological decisions and the total number of paths considered. We see that the listed studies employ various tools to summarize the large number of generated results, including plots of outcome distributions (e.g., boxplots) and sensitivity analyses with respect to specific forks. The latter is carried out by computing the conditional average of outcomes when one of the steps is fixed. For example, in Figure 1, this would involve averaging all returns with financial firms included and comparing them to the average returns with financial firms excluded, or averaging all equally weighted portfolio returns and comparing them to the average value-weighted returns. This is done to evaluate whether a specific step systematically produces notably different outcomes on average.

Study	Forks	Paths	Outcomes	Reported results	SE	NSE
Mitton (2022)	10	1,024	t-stat	distributions	-	-
Beyer and Bauckloh (2024)	11	116,640	alpha, AR, t-stat	distributions, sensitivity	-	IQR
Fieberg et al. (2024)	10	20,736	alpha, AR, SR	distributions, sensitivity	MSD	SD
Menkveld et al. (2024)	7-9	12,384	microstructure (6)	distributions, sensitivity, tests	-	IQR
Soebhag et al. (2024)	11	2,048	SR	distributions, sensitivity	MSD	SD
Walter et al. (2024)	14	69,120	AR, t-stat	distributions, sensitivity, tests	MSD	IQR
Cakici et al. (2025a)	10	69,120	alpha, AR, SR	distribution, sensitivity, tests	MSD	SD
Cakici et al. (2025b)	9	19,440	alpha, SR, t-stat	distributions, sensitivity	-	-
Chen et al. (2025)	9	1,056	AR	distributions, sensitivity	MSD	SD
Cirulli et al. (2025)	7	9,720	SR	distributions, sensitivity	MSD	SD

Table 1: **Multi-design studies in finance**. This table displays published articles and working papers in finance that explicitly consider a large number of methodological choices or *Forks* and a large number of *Paths*. *Outcomes* can be coefficients from regression models, t-statistics (t-stat), confidence intervals (CI), average returns (AR), Sharpe ratios (SR), or intercepts from factor models (alpha). For the nonstandard errors (*NSE*), IQR denotes the interquartile range of outcomes and SD is their standard deviation. Standard errors (*SE*) are taken to be the mean of standard deviations of outcomes (MSD), often obtained by bootstrapping returns of portfolios after to spanning the paths. In *Reported results*, sensitivity refers to analyses that investigate the impact of decisions and forks, while distributions encompass boxplots, densities, empirical cumulative distribution functions or particular summary statistics of outcomes. Tests mostly pertain to hypotheses on the existence of NSE and on the significance of variations across paths or forks.

Moreover, since the pioneering work of Menkveld et al. (2024), it has become common practice to report the so-called *nonstandard errors*. The latter aim to capture the impact of analysts' ad hoc methodological choices and is measured by taking either the interquartile range of outcomes or their cross-sectional standard deviation (see NSE column in Table 1). Menkveld et al. (2024) and Walter et al. (2024) are the only two studies formally testing whether nonstandard errors are statistically significant or not. This is done by testing whether individual-path outcomes differ from the overall median across paths. With re-

gard to the evaluation of SE, all the papers listed in Table 1 follow the same methodology. Indeed, for each anomaly and for each path, they generate a time series of returns. Then, they employ bootstrapping techniques on each series of anomaly returns to generate new averages and compute the corresponding standard deviation across samples. Finally, the SE is defined as the mean of these standard deviations across all paths. In contrast, we propose in this paper to use resampling *before* running the paths.

2.2 An example in asset pricing

To further motivate and illustrate our study, we review common variations in protocols in the asset pricing literature. Building on Chen and Zimmermann (2022a), we provide a brief overview of the choices made in the most influential papers in the field. As of January 2025, there are 331 anomalies in the dataset of the Open Source Asset Pricing project. Common decisions concern:[3]

- **Ad hoc filters**: Excluding certain regulated sectors (e.g., banks, real estate investment trusts, utilities).
- **Size filters**: Whether or not very small stocks are removed (e.g., bottom 5% or 10%), or based on the absolute price value (e.g., to exclude penny stocks). There is no common practice and the authors list 17 strategies used in the literature.
- **Imputation**: Whether missing data handling is performed cross-sectionally (using the mean or median), or chronologically (using the latest known value), or not performed at all.
- **Long-short quantile**: The sorting threshold that decides where to go long versus short. A majority of papers use quintiles (66 instances), deciles (61), but some authors also use quantiles at the 0.25, 0.30, or 0.50 levels.
- **Sample period**: The starting month for accounting data is most often taken to be June (190 instances) or December (54).
- **Stock weight**: The weighting scheme applied to sorted securities. Chen and Zimmermann (2022a) list 210 instances of equally-weighting, 32 of value-weighting, and 90 where this information is not disclosed.
- **Holding period**: How long the long-short portfolio is held before rebalancing. Monthly (120 instances) and annual periods (110) are by far the most common choices. Other options include quarterly (7) and biannual rebalancing (3).

This brief overview shows that the few commonly used options in the literature result in a wide array of choices. As an illustration, we see in Figure 1 that seven decisions lead to 576 possible paths. The full details of the construction of the long-short portfolios are postponed to Appendix B. We highlight two paths (the blue and orange ones) which have zero steps in common. Note that two different paths may share up to six steps in common.

[3]Other possible choices include leverage (Cirulli et al. (2025)), lookback window (Cirulli et al. (2025), Walter et al. (2024)), exclusion of sectors (Beyer and Bauckloh (2024), Soebhag et al. (2024), Walter et al. (2024)) or stocks with insufficient data (Walter et al. (2024)), multiple sorting (Beyer and Bauckloh (2024), Soebhag et al. (2024), Walter et al. (2024)), industry neutralization (Soebhag et al. (2024)), and alternative data vendors (Beyer and Bauckloh (2024)).

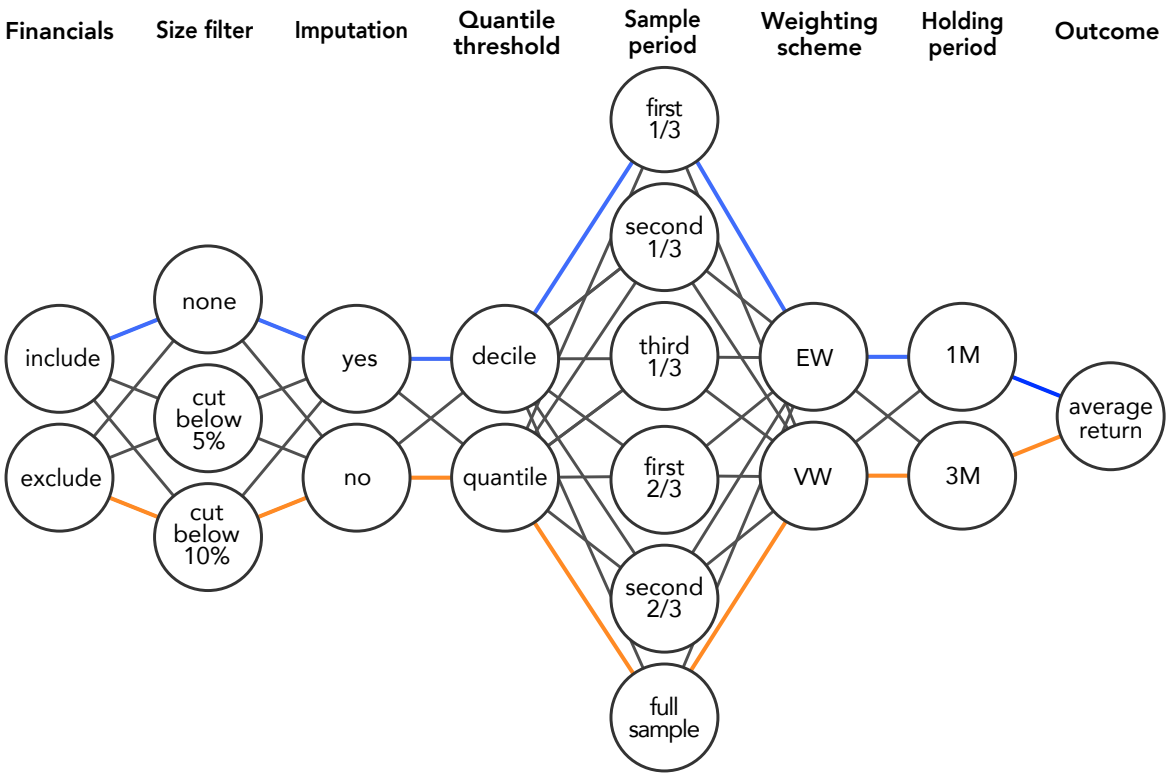

Figure 1: **Forking paths**. This figure displays the seven steps of the protocol along with the associated 576 paths. The ones in blue and orange follow entirely different steps.

2.3 Notations and definitions

We assume that the empirical part of any research process starts with a given dataset, which we call $\mathbb{D}$. A particular study is then modeled as a sequence of J operations f_j that occur successively. Formally, the reference research output $\hat{b}$ is given by:

$$\hat{b} := \hat{b}(\mathbb{D}) = f_J \circ f_{J-1} \circ \cdots \circ f_1(\mathbb{D}). \tag{1}$$

We assume that $\hat{b}$ is a scalar (e.g., an estimate, a t-statistic, or a p-value), but it may also be a more complex object, such as a vector (e.g., a confidence interval).

As an illustration, in Figure 1, there are $J = 7$ steps and the first one (f_1) pertains to whether or not include financial companies in the analysis, while the second one (f_2) filters out the smallest firms. Henceforth, we assume that each step f_j offers r_j deterministic options from which the researcher must choose, denoted by $\mathbb{f}_{j,r}$ for $r = 1, \ldots, r_j$, with $r_j \geqslant 2$. In Figure 1, $r_1 = 2$ (include or exclude financials) and $r_2 = 3$ (no screening plus two thresholds for the size filter). Visually, a path corresponds to a complete trajectory from left to right. The total number of paths is $P = \prod_{j=1}^{J} r_j$.

As any output is always associated with a given path, we use path indices: $\hat{b}_p$. Each output $\hat{b}_p$ is a random variable that depends on the realization of $\mathbb{D}$ as well as on the choice

of path p. This notation allows us to introduce the core concept of the paper, which is the correlation between the outcomes produced by two alternative paths p and q:

$$\rho_{p,q} = \mathbb{C}\text{or}\left(\hat{b}_p(\mathbb{D}), \hat{b}_q(\mathbb{D})\right). \tag{2}$$

Empirically, the above correlation is estimated through variations in the dataset $\mathbb{D}$. By resampling the dataset N times, we obtain N realizations of $\hat{b}_p(\mathbb{D})$ and $\hat{b}_q(\mathbb{D})$, which allows us to compute the sample correlation between the two series. Intuitively, $\rho_{p,q}$ measures how similar two paths are, with more overlap leading to higher correlation and less commonality leading to lower correlation.

3 The pernicious effect of correlated outcomes

3.1 Impact on the distribution of the effect

The goal of multi-design analyses is to produce multiple estimates in order to build a robust body of evidence regarding the effect of interest. A natural approach is to summarize the resulting estimates using means, medians, boxplots, etc. Such an approach relies on the assumption that the empirical distribution generated from these estimates closely approximates the true distribution of the effect. To assess whether this assumption holds in practice, it is necessary to evaluate the distance between the true (unknown) cumulative distribution function and its estimate from the sampled paths. Notably, Theorem 1 in Azriel and Schwartzman (2015) provides an upper bound on this distance. It states that if the effects are assumed to follow a standard multivariate Gaussian law with correlation matrix $\mathbf{\Sigma}_P = [\rho_{p,q}]_{1\leqslant p,q\leqslant P}$, then:

$$\sup_{x\in\mathbb{R}} \mathbb{E}\left[(\Phi(x) - \Phi_{\hat{b},P}(x))^2\right] \leqslant \frac{1}{4P} + C\|\mathbf{\Sigma}_P\|_1, \tag{3}$$

where $C > 0$ is some constant, which can be taken to be equal to $C = 1/2$ for simplicity, and the norm is $\|\mathbf{\Sigma}_P\|_1 = P^{-2}\sum_{1\leqslant p,q\leqslant P}|\rho_{p,q}|$.[4] The above result simply states that the error that one makes when confusing the empirical and true cumulative distribution functions is bounded from above by C times the average of the absolute correlations between paths. Indeed, this second term is in practice much larger than the first one ($1/4P$).

To illustrate how correlations affect the shape of the cumulative distribution function of the effect, we consider a stylized example involving a group of financial analysts. We assume that (1) each analyst provides one-year-ahead stock price forecasts for a sample of firms, and (2) the analysts are allowed to confer before submitting their individual forecasts, which introduces the possibility of persuasion effects and correlation among forecasts. Furthermore, we consider another group of financial analysts covering the exact same firms, but forming their predictions without any prior discussion. As a result, forecasts in the former group (group H for high) are more likely to be highly correlated than those in the latter group (group L for low).

[4] In the case of equal correlation among all paths, $\|\mathbf{\Sigma}_P\|_1 = \frac{P+\rho P(P-1)}{P^2}$. We see that the upper bound in (3) does not shrink to zero as P increases, unless $\rho = 0$.

In each group, we model the distribution of the forecast price variations across analysts using a multivariate Gaussian distribution with zero means and a group-specific correlation matrix.[5] As shown in the left panel of Figure 2, our simulation framework produces two plausible distributions of (non-diagonal) correlations, both displaying a large proportion of small correlations. In group H, 64% of the correlations across analysts are below 0.3 whereas this fraction is 81% in group L. Note that we allow for positive correlations in group L to account for the fact that analysts may process similar information or use the same pricing models. Importantly, under this stylized model, there is a 50% probability for each analyst to make a positive or negative price change.

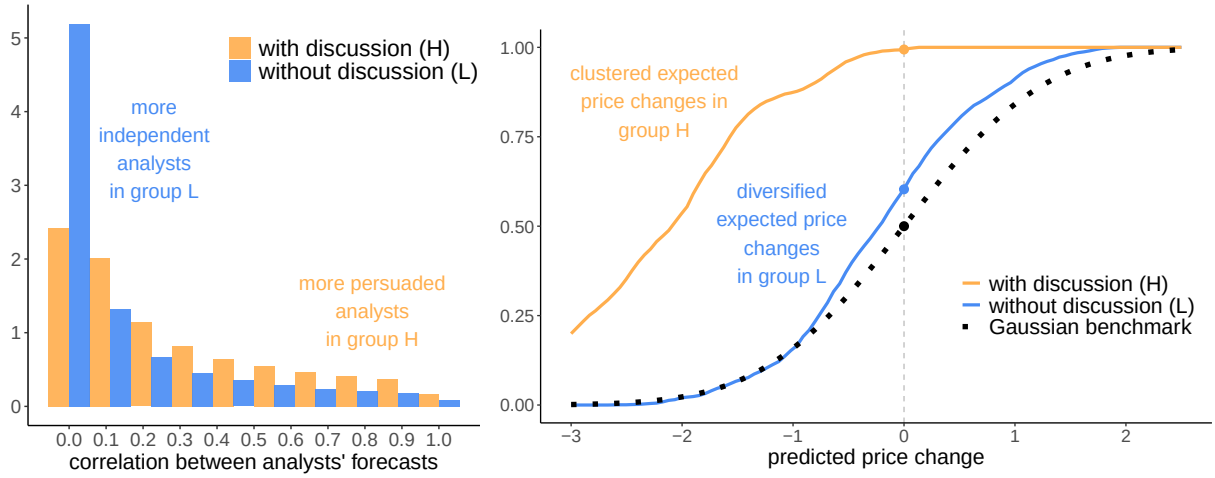

Figure 2: **Simulation exercise**. In the left panel, we plot the distributions of correlations for group H (with higher correlations due to discussions, in **orange**) and group L (with lower correlations, in **blue**). The simulation is based on Toeplitz matrices with ascending and descending diagonals equal to γ^n with $n = 0$ being the diagonal and $\gamma = 0.996$ for group L (yielding $\|\mathbf{\Sigma}_P\|_1 = 0.1525$) and $\gamma = 0.998$ for group H (corresponding to $\|\mathbf{\Sigma}_P\|_1 = 0.2776$). We set the number of analysts in each group to $P = 3,000$, which is also the number of rows/columns of the correlation matrix. In the right panel, we plot the cumulative distribution function of a single draw (i.e., firm) of the P expected price changes. A value of -1 on the x-axis represents a price change of -\$1. The black points mark the Gaussian density.

In the right panel of Figure 2, we display the cumulative distribution function of the predicted price change for a randomly selected firm. We see that the proportion of forecasts below zero is equal to 60% in the low-correlation case vs. 99% in the high-correlation case. This means that, in group H, discussions among analysts have led to a crowding towards bearish forecasts, as 99% of them predict a decline in the stock price. In group L, only 60% of analysts also foresee a contraction in the price.

As the true proportion of negative predictions is 50% for each individual (unbiased)

[5]To generate correlations and make sure they are higher in group H than in group L, we resort to a Toeplitz matrix that produces a distribution of positive correlations. The ascending and descending diagonals are equal to γ^n where $n = 0$ on the diagonal, $n = 1$ on the first superdiagonal, etc. In group H, we set γ to 0.998 which corresponds to $|\mathbf{\Sigma}_P\|_1 = 0.2776$ whereas in group L, $\gamma = 0.996$, which corresponds to $|\mathbf{\Sigma}_P\|_1 = 0.1525$.

analyst, the absolute error in the proportion is equal to $99\% - 50\% = 49\%$ in group H, but to only 10% in group L. These values are realizations of the error $\Phi(x) - \Phi_{\hat{b},P}(x)$ in Equation (3). Other random draws, corresponding to other firms, could yield smaller discrepancies. However, according to Equation (3), repeating this process a large number of times would lead to an average squared error that is smaller than $1/(4P)$ plus C times the values displayed in footnote (5).

This simulation exercise shows that even with moderate correlations between outcomes, the distance separating the empirical distribution and the true one can be substantial for any given draw.[6] However, an important question remains: how large are the correlations between outcomes in practice? To start answering this question, we conduct a preliminary analysis based on four popular asset pricing anomalies: the market capitalization (*size* factor), the book-to-market ratio (*value* factor), the past 12 month return (*momentum* factor), and the asset growth (*investment* factor).

To conduct our tests, we extract data from Chen and Zimmermann (2022a)'s website for the period September 1950 to December 2022. The baseline dataset comprises 4.475 million observations, although some rows contain missing values when certain characteristics are unavailable. The data are structured as an unbalanced panel: each row corresponds to a firm-month pair, while each column reports a characteristic (e.g., stock price, market capitalization, book-to-market, etc.), with the first two columns denoting date and firm. Further details on the anomalies are provided in Appendix A.

For each sorting characteristic, we resample the initial dataset by extracting sub-samples of the original data. Specifically, we randomly select rows of the data (i.e., month-stock pairs) that correspond to 40% of the initial dataset, without replacement.[7] Then, for each new sample, we run all 576 paths depicted in Figure 1. Since we use the same sample for all the paths, we denote this approach the *common resampling* approach. As we repeat this process $N = 500$ times, we end up for each factor with $500 \times 576 = 288,000$ estimates for the average return of the long-minus-short sorted portfolios. For each factor, we then compute all the correlations $\rho_{p,q}$ that populate the matrix $\mathbf{\Sigma}_P$.

Figure 3 displays the distribution of estimated off-diagonal correlations of outcomes across paths for the four sorting indicators. These distributions indicate that the correlations are substantial, with a pronounced concentration around 0.5. The norms of the corresponding correlation matrices, $\|\mathbf{\Sigma}_P\|_1$, are gathered in Panel A of Table 2. We note that all of these values are substantially higher (0.27-0.37) than those observed in our simulation exercise (0.15-0.28). Taken together, these results underscore the potentially large errors that may arise when using the empirical distribution of effects (here, the average returns of sorted portfolios) as a proxy for the true distribution.

[6] In this simulation, each draw represents a firm, and throughout the remainder of the paper, one draw denotes a sample.

[7] Depending on the empirical context (sample sizes, correlations in the data, etc.), one could resort to more advanced resampling techniques. We discuss bootstrapping techniques in Section 6.2.

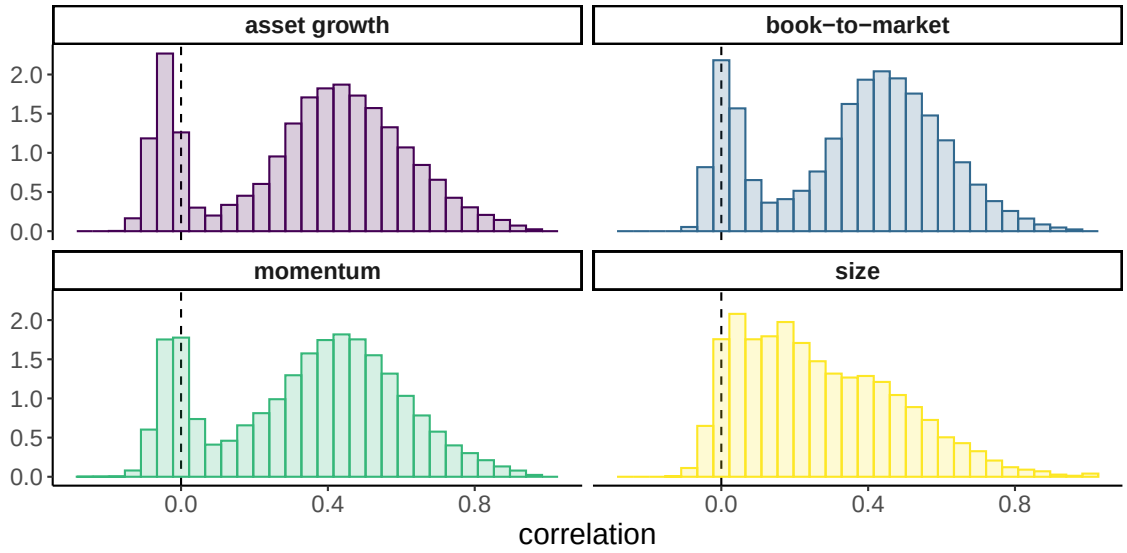

Figure 3: **Distribution of correlations**. We show the distribution of the correlations $\mathbb{C}or(\hat{b}_p, \hat{b}_q)$ for each of the four sorting variables, distinguished by color. Correlations are computed on 500 random subsamples with a number of rows equal to 40% of the original dataset.

	Factors			
PANEL A: Common resampling	Asset growth	Book-to-market	Momentum	Size
$P^{-2}\sum_{p,q}\lvert\hat{\rho}_{p,q}\rvert = \lVert\hat{\mathbf{\Sigma}}_P\rVert_1$	0.3647	0.3652	0.3557	0.2703
$P^{-2}\sum_{p,q}\hat{\rho}_{p,q}$	0.3448	0.3613	0.3430	0.2665
PANEL B: Specific resampling				
$P^{-2}\sum_{p,q}\lvert\hat{\rho}_{p,q}\rvert = \lVert\hat{\mathbf{\Sigma}}_P\rVert_1$	0.0377	0.0376	0.0376	0.0377
$P^{-2}\sum_{p,q}\hat{\rho}_{p,q}$	0.0021	0.0024	0.0020	0.0020

Table 2: **Sums of correlations**. We report the sum of absolute correlations used to compute the upper bound in Equation (3) and the sum of correlations used to compute the variance in Equation (6). Both are estimated from $N = 500$ samples for each of the four asset pricing anomalies. In Panel A, paths are run after *common resampling* (i.e., all paths use the same common dataset). In Panel B, paths are run after *specific resampling* (i.e., each path uses its own specific dataset).

3.2 Impact on the variance of the sample mean

We now provide a second illustration of the detrimental impact of large and asymmetric correlations. We start by showing how this translates into statistical testing for the mean. To construct confidence intervals for the mean μ_b, we define the sample mean estimator:

$$\hat{\mu}_b = \frac{1}{P}\sum_{p=1}^{P}\hat{b}_p. \tag{4}$$

and its variance:

$$\sigma^2_{\hat{\mu}_b} = \mathbb{V}\left[\hat{\mu}_b\right] = \frac{1}{P^2} \sum_{1 \leqslant p,q \leqslant P} \mathbb{E}\left[\left(\hat{b}_p - \mu_b\right)\left(\hat{b}_q - \mu_b\right)\right] = \frac{\sigma^2_b}{P^2} \sum_{p,q} \rho_{p,q} \tag{5}$$

$$= \underbrace{\frac{\sigma^2_b}{P^2}}_{\text{variance}} + \underbrace{\frac{\sigma^2_b}{P^2} \sum_{p,q} \rho_{p \neq q}}_{\text{covariance}}, \tag{6}$$

where $\sigma^2_b = \mathbb{V}[\hat{b}_p] = \mathbb{V}[\hat{b}_q]$. This identity leads to two important observations. First, the construction of confidence intervals for μ_b requires information on the uncertainty of $\hat{\mu}_b$, which is captured by the estimation of $\sigma^2_{\hat{\mu}_b}$. As such, the intervals will depend on the correlations between paths, as we will show in the next section.

Second, the intervals will be tighter and more informative if these correlations are small and/or symmetric around zero. The fact that variance reduction can come from lower correlations is well documented in other areas. For instance, in forecasting, combining models with uncorrelated errors yields the best results, but estimating correlations across models is hard (Timmermann (2006), Wang et al. (2023)). Similarly, in machine learning, bagging performs best when aggregating predictions with low correlations (Breiman (2001)).

As shown in Equation (6), the variance of the sample mean $\sigma^2_{\hat{\mu}_b}$ is smaller than σ^2_b because all pairwise correlations $\rho_{p,q}$ are smaller than one. The extent of the difference between the two variances depends on the values of these correlations. If they are large and positive, then $\sigma^2_{\hat{\mu}_b}$ will be close to σ^2_b, and thus quite large. This is often the case, since highly similar paths tend to produce highly correlated outcomes (see Figure 3). In Panel A of Table 2, we see that the sums of correlations, which are the rightmost terms in Equation (6), range between 0.26 and 0.37. These high levels imply large variances for $\hat{\mu}_b$.

Importantly, the variance expression in Equation (6) bears a close connection to the heteroskedasticity and autocorrelation consistent (HAC) corrections introduced by Newey and West (1987). In contrast to the temporal dependence typically considered in that setting, the correlation here originates from cross-sectional dependence across paths.

We will show in Section 4 how path-specific resampling can mitigate correlation among outcomes and greatly improve inference.

3.3 The origins of correlations

Where do these positive correlations come from? Part of the answer lies in the commonality between paths. Suppose that two researchers make exactly the same methodological choices, except for outlier management. Arguably, this minor variation would only lead to a small change in the outcome. Hence, because the paths are very similar, we can expect that their results will be highly correlated. Conversely, if researchers follow paths that rarely or never overlap, then the correlation should be much lower.

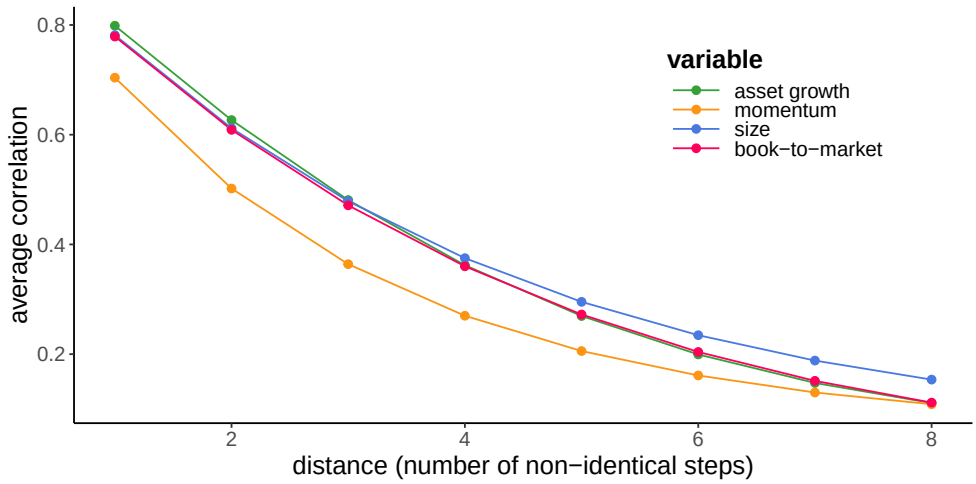

Figure 4: **Average correlation as a function of path distance**. We plot the mean correlation across all pairs of paths, as a function of the number of different choices between two paths (the distance between paths).

As this is a testable assumption, we propose examining it empirically. To do so, we define $d(p, q)$ as the number of choices that differ between path p and path q.[8] In Figure 4, we plot the average correlation between outcomes as a function of the path distance $d(p, q)$. By definition, $d(p, q)$ lies between one (for $p \neq q$) and the number of steps that the researcher can make in the protocol (J). We observe a power decay: as the distance increases, the average correlation decreases. It is reasonable to expect that, if the number of steps is arbitrarily large, the correlations between two distant paths would approach zero.

4 The path-specific resampling strategy

4.1 The basic idea

The above analysis indicates that multi-design studies relying on a single version of a dataset offer limited guarantees for statistical inference. Indeed, as the numerous analyses carried out are likely to be highly correlated, we expect significant differences between the empirical and the true distributions. In this section, we show that specific resampling, which consists of generating a new dataset before running each path, is a simple and efficient way to mitigate this problem.

We start by showing that resampling the data before running the paths significantly alters the distribution of correlations. In order to clarify the difference between the two sampling schemes, we outline their steps in Table 3. As in Figure 3, we randomly select 40% of the original sample, but we do so *before* running any path. This approach shares

[8] Formally, $d(p, q) = \#\{j, r_{p,j} \neq r_{q,j}\} \in \{0, 1, \ldots, J\}$, where the operator $\#\{A\}$ measures the size (number of elements) of set A and $r_{p,j}$ is the option through which path p passes for step j.

strong similarities with subsampling in ensemble methods.[9] By picking only 40% of the original sample, we hope to strongly curtail the correlations between path outcomes and thus trim the variance defined in Equation (6).

common resampling	**path-specific resampling**

given a dataset $\mathbb{D}$ and a set of paths:
for bootstrap iteration $n = 1, \ldots, N$, **do**:
  generate new sample $\mathbb{D}_n$
  for path $p = 1, \ldots P$ **do**:
  generate outcome $\hat{b}_p(\mathbb{D}_n)$
	given a dataset $\mathbb{D}$ and a set of paths:
	for bootstrap iteration $n = 1, \ldots, N$, **do**:
	  for path $p = 1, \ldots P$ **do**:
	  generate new sample $\mathbb{D}_{n,p}$
	  generate outcome $\hat{b}_p(\mathbb{D}_{n,p})$

Table 3: **Pseudo-algorithm of sampling procedures**. The only difference between the two methods is that the third and fourth lines are swapped.

Figure 5 displays the empirical density obtained by subsampling the data before each new iteration of the protocol. We clearly see that the estimated correlations are symmetric around zero and highly concentrated within the [-0.1,0.1] interval. Furthermore, the distributions are very similar across all four characteristics. Coincidentally, this is approximately the distribution one would obtain when estimating correlations among independent Gaussian variables, based on a sample of 500 observations. This tends to suggest that the outcomes are in fact independent but that the estimation of correlations is noisy. As shown in Panel B of Table 2, the norm of the correlation matrix in this case is close to 0.04 for all four factors, which is almost a tenfold reduction compared to the case where we use the same dataset for all the paths (see Panel A of Table 2).

4.2 Inference on the mean

In this section, we lay out ways to carry out inference on path-generated outcomes. We are first interested in the empirical mean effect, $\hat{\mu}_b$ defined in Equation (4), which will naturally serve as a proxy for the mean μ_b. In order to proceed with inference, we must make some hypotheses about the underlying effect and how it is characterized by paths. We lay out a mild assumption on which we will rely for the remainder of the paper.

Assumption 1. *The estimated effects $\hat{b}_p$ are identically distributed and their law is (i) unimodal, (ii) symmetric around its mode and mean μ_b, and (iii) has a finite variance, σ_b^2.*

Henceforth, we seek to build confidence intervals of the following form:

$$\text{CI}_\alpha = [\hat{\mu}_b - \Delta_\alpha, \hat{\mu}_b + \Delta_\alpha], \tag{7}$$

and for which the probability that the true mean μ_b belongs to this interval is at least $1 - \alpha$, e.g., with $\alpha = 0.05$:

$$\mathbb{P}[|\mu_b - \hat{\mu}_b| \leqslant \Delta_\alpha] \geqslant 1 - \alpha. \tag{8}$$

[9] See, e.g., Zaman and Hirose (2009), Bühlmann (2011), and Duroux and Scornet (2018).

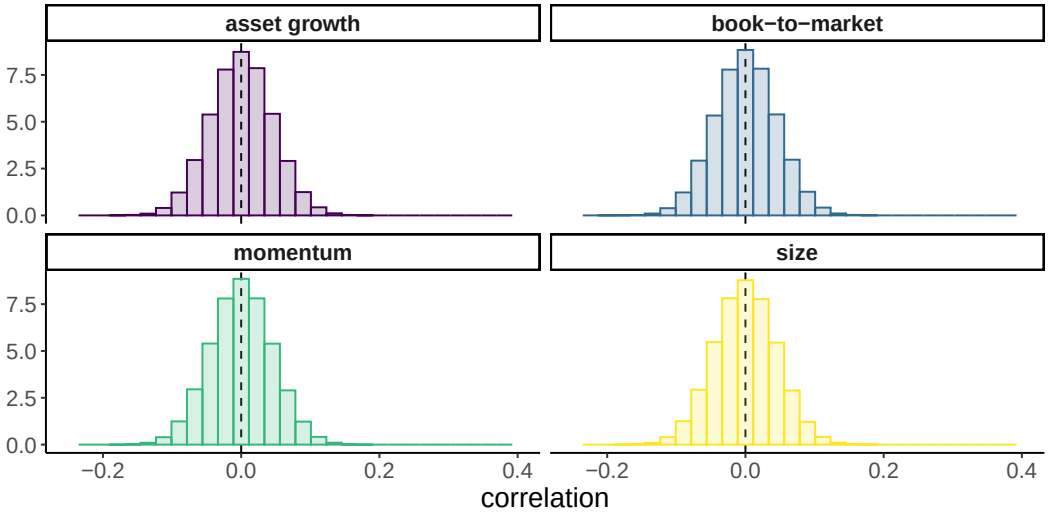

Figure 5: **Distribution of correlations after resampling**. We show the distribution of the correlations $\mathbb{C}or(\hat{b}_p, \hat{b}_q)$ for each of the four sorting variables, coded with colors. In this case, the data are subsampled before running each path and the number of rows corresponds to 40% of the size of the initial sample, picked without replacement. Correlations are computed on 500 random subsamples.

To set the value of Δ_α, which defines the width of the interval, we use the following Bienaymé-Chebyschev inequality, taken from Theorem 6.2 in Ion et al. (2023).[10]

$$\mathbb{P}[|Z - \mathbb{E}[Z]| \leqslant v] \geqslant 1 - \left(\frac{2\sigma_Z}{3v}\right)^2 \tag{9}$$

for $v \geqslant 2\sigma_Z/\sqrt{3}$. Setting $\alpha = (2\sigma_Z/(3v))^2$ and $Z = \hat{\mu}_b$ yields:

$$\mathbb{P}\left[|\hat{\mu}_b - \mu_b| \leqslant \frac{2\sigma_{\hat{\mu}_b}}{3\sqrt{\alpha}}\right] \geqslant 1 - \alpha, \tag{10}$$

where $1 - \alpha \in (0, 1]$ is the targeted level of confidence and $\sigma_{\hat{\mu}_b}$ is the standard deviation of $\hat{\mu}_b$ defined in Equation (6). Intuitively, when $\sigma_{\hat{\mu}_b}$ decreases, the confidence interval shrinks, making it more informative. Conversely, as α diminishes to increase the confidence level, the interval widens.

The only remaining challenge is the calculation of $\sigma_{\hat{\mu}_b}$. The crux of the problem lies in the estimation of the correlations, $\hat{\rho}_{p,q}$, based on N samples, which generates some additional uncertainty due to estimation error. We postpone the technical discussion of this

[10]Other concentration inequalities could be used. Those that consider sums of variables are however hard to apply. Indeed, the most general conditions for dependence are provided in Jirak (2023) and they assume a natural ordering of variables for which is it possible to define the speed at which memory between variables (e.g., autocorrelation) fades. The problem here is that, in presence of multiple paths, there is no such natural ordering. Therefore, we cannot resort to these inequalities. One could nevertheless resort to simpler concentration inequalities, such as Hoeffding (1963)'s bound, but treating the average as a single random variable. The main issue in this case is the derivation of the support of the random variable, which is far from obvious in the case of the sample mean.

point to Appendix C, and we show that, with probability at least $1-\alpha$, the following upper bound holds for $\sigma^2_{\hat{\mu}_b}$:

$$\sigma^2_{\hat{\mu}_b} \leqslant \overbrace{\frac{2\kappa\sigma^2_*}{3\sqrt{\alpha N}}}^{\text{penalty from estimation error}} + \frac{\sigma^2_*}{P^2}\sum_{p,q}\hat{\rho}_{p,q}, \tag{11}$$

where κ is a constant that depends on the correlations $\rho_{p,q}$ (see Appendix C) and σ^2_* is an upper bound for σ^2_b (see Appendix D). Overall, our result relies on confidence intervals for three quantities: $\hat{\mu}_b$, the primary quantity of interest, its standard deviation $\sigma_{\hat{\mu}_b}$, and the average of correlations, $\sum_{p,q}\frac{\rho_{p,q}}{P^2}$, and we need these intervals to hold *jointly*. Because it is extremely complex to properly model the dependence between the three terms, we conservatively assume that they are independent. As shown in Appendix C, this requires dividing α by three. Plugging the above expression into Equation (10), we obtain the final expression for the width of the confidence interval (7):

$$\Delta_\alpha = \frac{2\sqrt{3}\sigma_{\hat{\mu}_b}}{\sqrt{\alpha}} \leqslant \frac{2\sqrt{3}\sigma_*}{\sqrt{\alpha}}\sqrt{\frac{2\sqrt{3}\kappa}{\sqrt{\alpha N}} + \sum_{p,q}\frac{\hat{\rho}_{p,q}}{P^2}}. \tag{12}$$

As an illustration, we plot in Figure 6 the confidence intervals for the four asset pricing anomalies using the rightmost term of Equation (12). We consider two cases. First, we suppose that correlations are evaluated from paths run on identical samples (*common resampling*, displayed in yellow), which corresponds to the distributions shown in Figure 3. The second case pertains to situations where new samples are drawn prior to running a given path (*specific resampling*, displayed in blue), hence the distributions of correlations are those depicted in Figure 5. The non-diagonal $\hat{\rho}_{p,q}$ (i.e., $p \neq q$) are then such that their sum is negligible, hence the double sum in Equation (12) boils down to $P/P^2 = 1/P$ due to the diagonal terms.

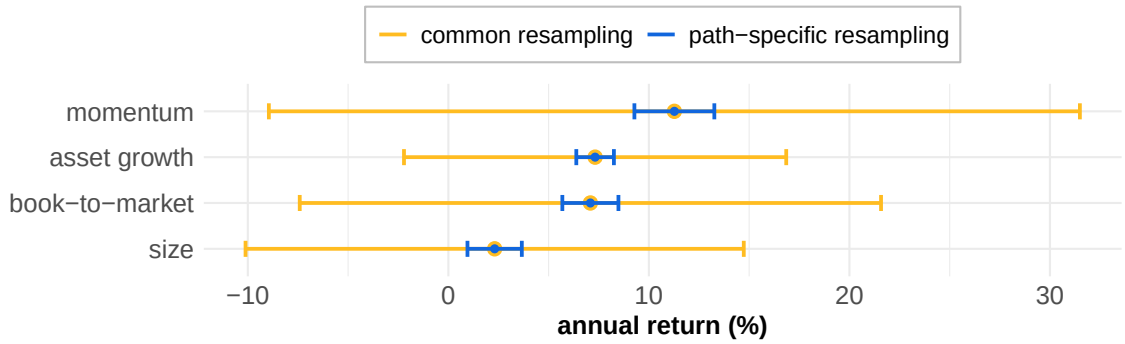

Figure 6: **Inference on the mean**. We plot, for $\alpha = 0.05$, the confidence intervals (7) of the mean of long-short portfolio returns in two situations: (i) common resampling corresponding to Figure 3 in **yellow** and (ii) path-specific resampling in **blue**. By definition, the sample means lie in the middle of the intervals. The width of intervals is given by Δ_α in Equation (12).

In the empirical derivations, σ_* is calculated as follows. First, for each of the $N = 500$ samples, we compute the standard deviation of effects across paths, which yields 500 estimates of σ_b^2. Next, to be as conservative as possible, we take the maximum of these values. Finally, we then use the upper bound σ_*^2 defined in Equation (34) in Appendix D.

Clearly, the ranges of the intervals in Figure 6 illustrate the clear benefits of path-specific resampling for inference on the mean. The width of the yellow interval is roughly ten times larger than that of the blue one. The order of magnitude of this difference was to be expected, given the figures in Table 2. Indeed, a sum of correlations that is more than 100 times larger is expected to yield, via Equation (12), an interval at least ten times wider.

4.3 Variance decomposition: standard vs. nonstandard errors

Given our focus on inference, the most critical quantity in this paper is the variance of the mean effect:

$$\sigma_{\hat{\mu}_b}^2 = \frac{\sigma_b^2}{P^2} \sum_{p,q} \rho_{p,q}. \tag{13}$$

Until now, we have focused on the rightmost component of this variance, namely the correlation coefficients. In this section, we analyze the other important term, σ_b^2, the variance of the effect, and we show how it relates to standard and nonstandard errors.

The standard error of an estimate pertains to uncertainty related to sampling. In the asset pricing literature, it is often estimated via bootstrapping returns, but, as shown in the review by Horowitz (2019), there are many ways to carry this out (e.g., parametric versus non-parametric methods, with or without blocks, etc.). To estimate the standard deviation, the contributions listed in Table 1 resample outcomes after spanning the paths.

By contrast, the nonstandard error of a given result refers to dispersion in outcomes across multiple paths. Two definitions are currently used in the literature: the interquartile range (Menkveld et al. (2024); Walter et al. (2024)) and the standard deviation (Fieberg et al. (2024); Soebhag et al. (2024)) of the cross-section of outcomes.

The current approaches have two shortcomings. First, the variety of estimation techniques leads to many different SE (and NSE) estimates, and we lack theoretical guidance for choosing among them. Moreover, the results vary across approaches, as shown in the right panel of Figure 7. Second, the lack of integration between SE and NSE estimation prevents these components from summing to the total variance of the mean effect.

In what follows, we propose a new approach to jointly evaluate both the standard and nonstandard errors. This common estimation framework leads to a single decomposition of the variance of the effect, which complies with the additivity property.

We first recall our notation $\hat{b}_p(\mathbb{D})$ for estimated effects, where $\mathbb{D}$ represents the dataset sample and p denotes the path. Naturally, these outcomes exhibit variations, and it is crucial to determine their origin, whether they arise from random fluctuations in the dataset or from methodological choices. To distinguish the sources of uncertainty, we employ the

law of total variance, which we state below for two random variables X and Y with finite variance (Theorem 9.5.4 in Blitzstein and Hwang (2019)):

$$\mathbb{V}[Y] = \mathbb{V}[\mathbb{E}[Y|X]] + \mathbb{E}[\mathbb{V}[Y|X]]. \tag{14}$$

We aim to decompose the variance of the effect $\hat{b}_p(\mathbb{D})$, which can be done by conditioning either on samples or on paths. For instance, conditioning with respect to samples leads to:

$$\sigma_b^2 = \mathbb{V}\left[\hat{b}_p(\mathbb{D})\right] = \underbrace{\mathbb{V}\left[\underbrace{\mathbb{E}[\hat{b}_p(\mathbb{D})|\mathbb{D}]}_{\text{avg effect across paths}}\right]}_{\text{variance of avg effect across samples}} + \underbrace{\mathbb{E}\left[\underbrace{\mathbb{V}[\hat{b}_p(\mathbb{D})|\mathbb{D}]}_{\substack{\text{variance across paths} \\ \text{NSE reported in literature}}}\right]}_{\text{average variance across samples}} \tag{15}$$

In the above expression, we notice that, in the second term, we recover the NSE that corresponds to the variance across paths, but only for one given dataset $\mathbb{D}$. This version of the law of total variance shows that this variance should then be averaged across several $\mathbb{D}$. The first term, which corresponds to the variance of the averages across datasets $\mathbb{D}$ has, to the best of our knowledge, never been reported in the literature.

An alternative version of the law of total variance can be obtained by conditioning with respect to each path p. In this case, we obtain the following sum:

$$\sigma_b^2 = \mathbb{V}\left[\hat{b}_p(\mathbb{D})\right] = \underbrace{\mathbb{V}\left[\underbrace{\mathbb{E}[\hat{b}_p(\mathbb{D})|p]}_{\text{avg effect across samples}}\right]}_{\text{variance of avg effect across paths}} + \underbrace{\mathbb{E}\left[\underbrace{\mathbb{V}[\hat{b}_p(\mathbb{D})|p]}_{\text{variance across samples}}\right]}_{\substack{\text{average variance across paths} \\ \text{SE reported in literature}}} \tag{16}$$

In this expression, the second term corresponds to the standard errors reported in the literature featured in Table 1. The first term, however, is new. It first computes, for each path, the average effect across samples and then evaluates the variance across paths. When there is a unique dataset, this first term is equal to the NSE, defined as the variance of the outcomes across paths. In Equation (16), paths will contribute to the dispersion of outcomes via $\mathbb{V}[\mathbb{E}[\hat{b}|p]]$, while sampling matters through $\mathbb{E}[\mathbb{V}[\hat{b}|p]]$.

The two identities above highlight each dimension individually (paths and sampling), emphasizing that the SE and NSE reported in the literature are not directly comparable, as they arise from different variance decompositions. Since both equations (16) and (15) (i) provide valid decompositions of $\mathbb{V}[\hat{b}_p(\mathbb{D})]$, (ii) are readily computable within our framework, and (iii) have no inherent reason to be preferred over one another, we propose the

following unified definitions for standard and nonstandard errors:

$$\text{SE} = \sqrt{\frac{\mathbb{E}[\mathbb{V}[\hat{b}|p]] + \mathbb{V}[\mathbb{E}[\hat{b}|\mathbb{D}]]}{2}}, \tag{17}$$

$$\text{NSE} = \sqrt{\frac{\mathbb{V}[\mathbb{E}[\hat{b}|p]] + \mathbb{E}[\mathbb{V}[\hat{b}|\mathbb{D}]]}{2}}. \tag{18}$$

The finite sample expressions of the above quantities are rigorously defined in Appendix E. Crucially, in contrast with the conventions previously used in the literature, these identities verify:

$$\sigma_b^2 = \text{SE}^2 + \text{NSE}^2. \tag{19}$$

This decomposition expresses the total variance of the effect as the sum of two components: one arising from sampling variability, and the other from the variability across paths. Several similar ideas or identities have been proposed in other contexts. For instance, Abadie et al. (2020) also point out the duality between sampling and protocol variation and they refer to the average of the conditional variances but not to the variance of the conditional averages. Therefore, an exact decomposition of the total variance is not possible in their context. In a similar vein, Holzmeister et al. (2024) define heterogeneity in outcomes as the variation in effect size estimates over and above sampling variation.

It is reasonable to question how the SE defined in Equation (17) compares to those commonly reported in the literature (see Table 1). To shed light on this, we conduct the following analysis. For the four baseline anomalies, we consider all 576 paths illustrated in Figure 1, generating return series for each long-short portfolio sorted on the corresponding firm characteristic. As in Soebhag et al. (2024) and Fieberg et al. (2024), we then apply a bootstrap procedure, resampling the returns (with replacement) to match the original sample size and computing the average returns for each new sample. We compute for each path the standard deviation of the bootstrapped averages, and then take the average of these standard deviations across all paths. The left panel of Figure 7 presents the results of this procedure, alongside the SE values obtained from Equation (17). We see that (1) the outcomes are quite consistent across anomalies and that (2) our approach produces SE values that are approximately half the size of those typically reported in the literature.

One may also wonder how the NSE defined in (18) compares to the single sample approach used until now in the literature. In the right panel of Figure 7, we display the distribution of NSEs evaluated on single datasets, across datasets. While the range is limited, it still underlines that in this case, the NSE remains subject to uncertainty and should be averaged, as shown in the last term of Equation (16). In the right panel, we also report the NSE as defined in Equation (18) and it is always smaller than the single-sample values currently used in the literature. The reason for this is simply that the single-sample component (the second one in Equation (18)) is larger than the path-focused component. Hence, upon averaging, the composite NSE is slightly lower than the single-sample NSE. The relative importance of SE and NSE in Equation (19) will be further investigated in Section 5.2 below.

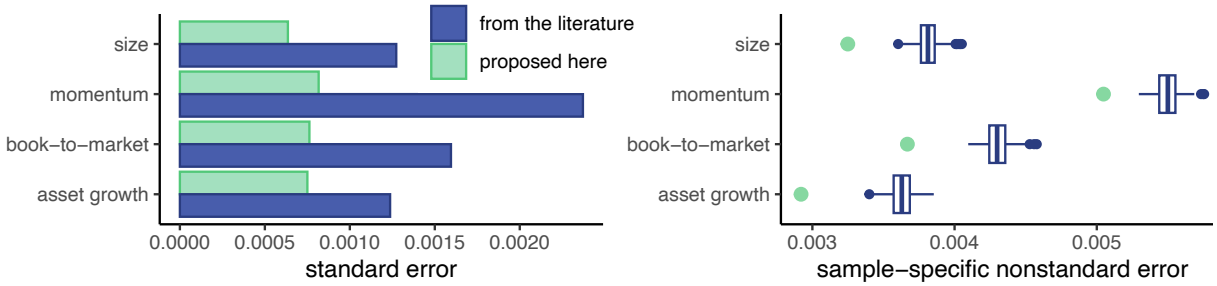

Figure 7: **Comparing SE and NSE methods**. In the left plot, we report the standard error for both our method (Equation (17)) and from the bootstrapping method suggested in the literature. In the right plot, we show the distribution of the NSE as it is computed in the literature (standard deviation of outcomes across paths for a single dataset), across the $N = 500$ samples we originally generated. In addition, the larger green points mark the NSE calculated as in Equation (18).

Finally, and most importantly, this analysis of the total variance allows us to simultaneously showcase the three components of uncertainty in multi-design studies. By plugging Equation (19) into Equation (13), we can rewrite the variance of the estimator of the mean:

$$\sigma^2_{\hat{\mu}_b} = \left(\text{SE}^2 + \text{NSE}^2\right) \sum_{p,q} \frac{\rho_{p,q}}{P^2}. \tag{20}$$

In the favorable case where correlations are approximately symmetric, such that the last term approaches $1/P$ (e.g., when resampling is performed independently for each path), the variance above depends primarily on SE, NSE, and the number of paths P. As suggested by the previous results, and later confirmed in the more extensive analysis below, the NSE tends to clearly dominate the SE. Consequently, methodological variation becomes the main determinant of the width of the confidence interval for the mean effect.

5 Uncovering persistent anomalies and NSE at scale

5.1 Resilient anomalies

We now turn to the application of the methods described above to a broader set of anomalies. We consider 33 factors out of the hundreds of factors reported in the literature. We rely on the Open Source Asset Pricing dataset of Chen and Zimmermann (2022a), and only keep the sorting variables that satisfy the following three criteria:

1. The variable is continuous and not discrete. Indeed, as we use several sorting thresholds in the paths, this is only suited for variables taking arbitrarily large numbers of values in the cross-section.
2. Data coverage is available for at least 500 stocks starting in 1950. This is because we sometimes use deciles for sorting, setting a minimum of 500 assets implies long and short legs of 50 stocks, which is the bare minimum to ensure diversification.

3. Finally, we wish to compare our results with those in the original published papers. This information is provided here by Chen and Zimmermann (2022a). Hence, the last requirement is that the average return be reported for the sorting variable.

In the end, intersection of these conditions leads to 33 factors. For each of them, we implement the 576 methodological paths presented in Figure 1. Each path leads to a time series of portfolio returns after the final weighting step and these returns are averaged to yield the performance of the factor. We repeat this analysis 500 times for each path and obtain a total of 288,000 estimated average returns of a long-short portfolio. We display the distribution of these estimates for each factor in Figure 8, along with the average return reported in the first academic study introducing this particular anomaly as reported by Chen and Zimmermann (2022a). In Figure 9, we report the confidence intervals defined by (12). We omit the intervals from common resampling, as this approach is clearly suboptimal (i.e., excessively wide) and would not be used for inference in this context.

Taken together, the two figures reveal a variety of situations. First, there are cases in which our results fully corroborate the original publications. This holds true, for instance, for the *Mom6m*, *IndMom*, and *BetaTailRisk* variables. Indeed, the original average returns fall in the middle of our intervals and none of them overlap with zero. We also find some rare cases for which our results are more favorable: this occurs when the original returns lie to the left of the intervals (*High52*, *CBOperProf*, *OperProfRD*, *GP*). There are also many occurrences in which the original results lie far to the right (McLean and Pontiff (2016)), but the latter are also to the right of zero, meaning that anomalies are nonetheless confirmed.

We also notice a variety of widths for the confidence intervals. This comes primarily from the cross-path dispersion of outcomes. Narrow intervals signal that variables can sustain a lot of methodological changes with limited changes in performance. However, large intervals indicate that factors are more sensitive to implementation choices. Finally, we also find instances of asset pricing factors that are not significant upon specific resampling, i.e., for which the interval encompasses zero.

5.2 Standard vs. nonstandard errors

We implement the variance decomposition outlined in Equations (17)-(19) for the 33 factors. Doing so allows us to contribute to the ongoing debate in the literature on the importance of variations due to differences in research design across researchers, i.e., NSE. In their multi-analyst study in the field of microstructure, Menkveld et al. (2024) characterize the NSE associated with their six types of estimates as "*sizable*". However, they do not compare them directly to the associated SE. Such direct comparison is made by Soebhag et al. (2024) in their analysis on the Sharpe ratios of sorted portfolios. They find that the magnitude of the NSE and SE are more or less comparable: for the ten factors they consider, they report that the NSE-to-SE ratio lies between 0.5 and 2. Nevertheless, as we argue in Section 4.3, SE and NSE can be meaningfully compared only when they are linked to a common reference quantity, which we propose should be the variance of the effect under investigation.

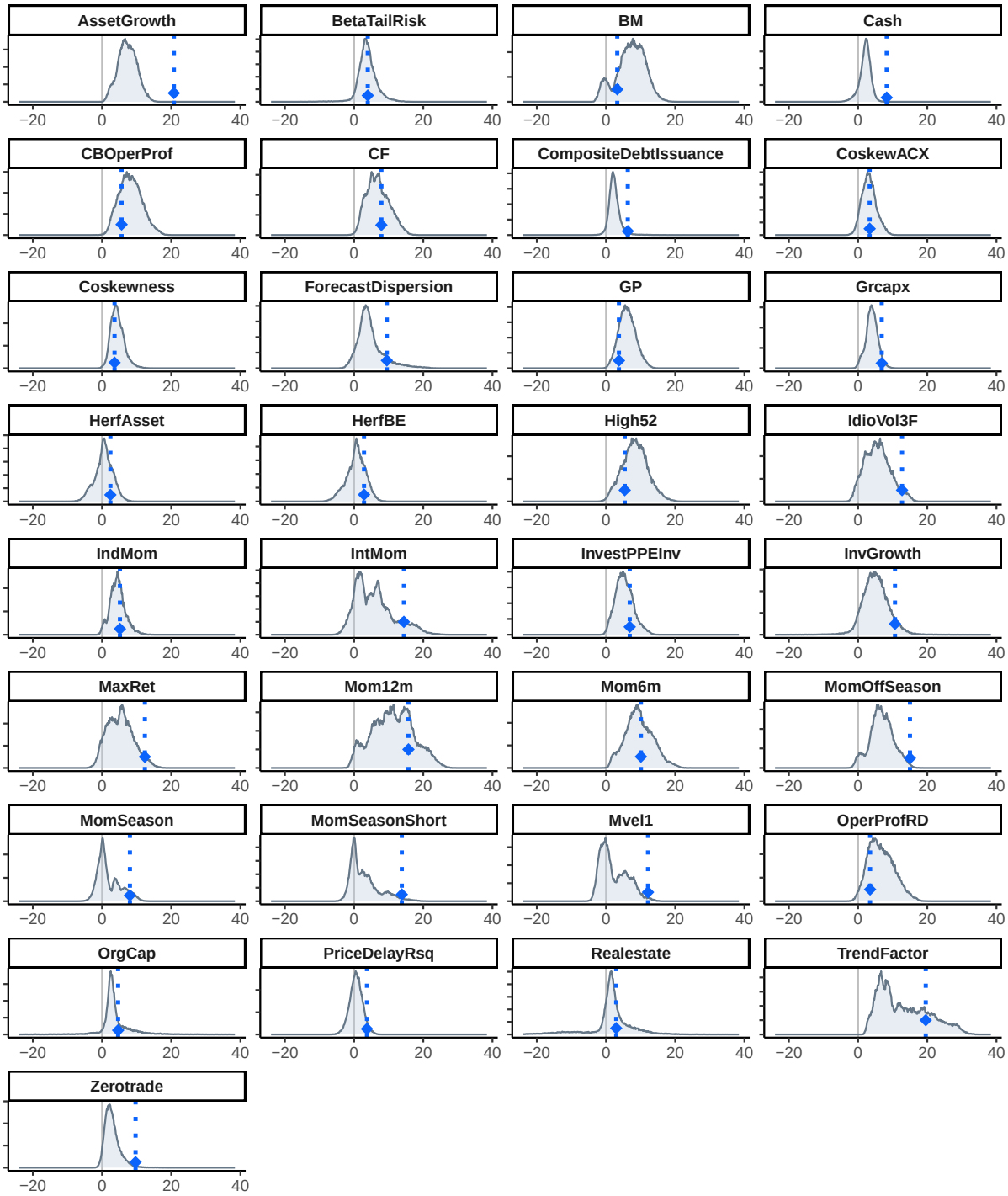

Figure 8: **Distribution of average returns**. We plot the distribution of average annual returns (in percents) across all 576 paths, bootstrapped samples, and sampling schemes. The blue diamonds and the vertical points mark the estimates first reported in the literature. The vertical gray line shows the zero return. The names of characteristics and the corresponding returns are those of Chen and Zimmermann (2022a), see Appendix A.

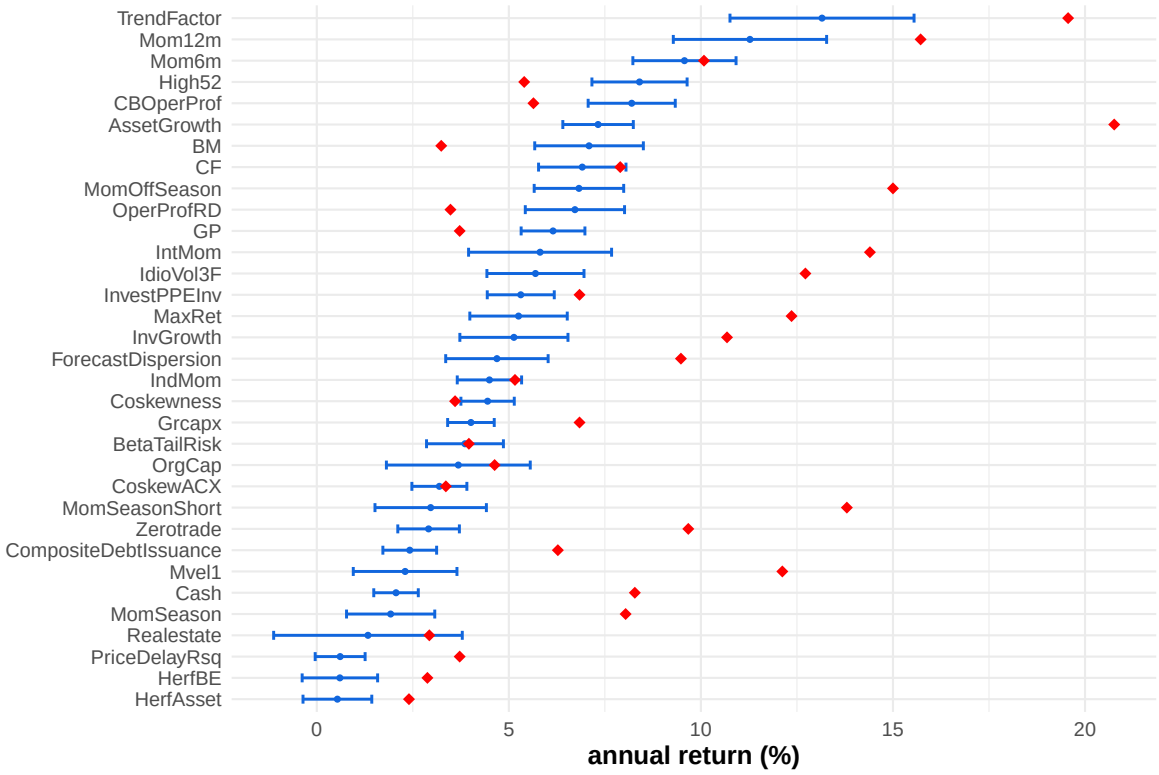

Figure 9: **Resilient anomalies**. We plot, for $\alpha = 0.05$, the confidence intervals (7) of the mean of long-short returns under path-specific resampling. By definition, the sample means lie in the middle of the intervals. The width of intervals is given by Δ_α in Equation (12). The **red** diamonds locate the average return in the original studies.

Our empirical evidence is in line with the findings reported by Menkveld et al. (2024). We show in Figure 10 that variations in methodologies have a strong impact on the final estimate, as shown by the large NSE reported in the figure. However, this effect varies across anomalies. The total variance of the most sensitive factor, *TrendFactor*, is more than ten times greater than that of the least sensitive one, *Cash*. Consequently, the widest intervals in Figure 9 are roughly three times broader than the narrowest, as their range scales with the square root of the effect's variance.

We also see that, for most anomalies, the NSE is at least ten times larger than the SE. In a handful of cases, it even exceeds 20 times the SE. This finding suggests a more important role for NSE than previously claimed in the literature (Soebhag et al. (2024)) and shows the importance of deriving both SE and NSE in a common framework.[11] Finally, we notice that the two anomalies (*Realestate* and *OrgCap*) whose standard errors are substantially larger than those of the other anomalies. Upon closer examination, this comes from the first component of the SE, namely $\mathbb{E}[\mathbb{V}[\hat{b}|p]]$ in Equation (19). It means that, on average, the

[11] Applying the methodology from the literature, we obtain an NSE-to-SE ratio between 2 and 3 (see Figure 7), which closely aligns with the findings of Soebhag et al. (2024).

path outcomes for these two characteristics are more sensitive to random sampling than those of the other anomalies. Still, the dispersion of their average returns across paths is similar to those of other factors.

Our results about the magnitude and relative importance of NSE have important implications in terms of inference for multi-design studies. Indeed, as shown in Section 4.2, the confidence interval for the empirical mean effect critically depends on its variance, which we derive as follows: $\sigma^2_{\hat{\mu}_b} = (\text{SE}^2 + \text{NSE}^2) \sum_{p,q} \rho_{p,q}/P^2$. As the SE term is dwarfed by the NSE term and the final correlation term is small with path-specific resampling (see Table 2), the main driver of the variance is the NSE. As a consequence, we claim that researchers need to internalize the uncertainty about methodological choices by default in any protocol whenever this is feasible.

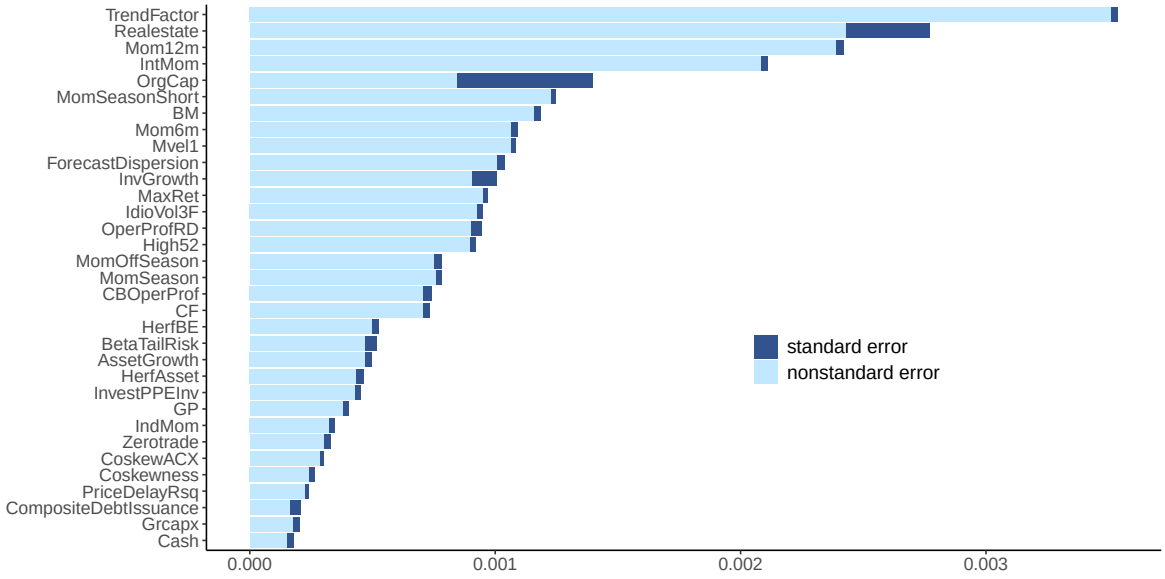

Figure 10: **Variance decomposition: standard versus nonstandard errors**. We display the decomposition of the variance in average returns proposed in Equations (17)-(19) and formally defined in Appendix E. The names of characteristics are those of Chen and Zimmermann (2022a).

6 Extensions

6.1 Differential weighting

Thus far, we have described an agnostic approach that treats all outcomes as equally important. However, alternative weighting schemes can be considered. One possibility is to assign greater weight to specific paths, effectively counting them multiple times, if they are more probable or if they are supported by stronger scientific reasoning. In this case,

the sample mean estimator is written:

$$\hat{\mu}_b = \sum_{p=1}^{P} \omega_p \hat{b}_p, \quad \text{with} \quad \sum_{p=1}^{P} \omega_p = 1. \tag{21}$$

and its variance:

$$\sigma^2_{\hat{\mu}_b} = \mathbb{V}\left[\hat{\mu}_b\right] = \sigma^2_b \sum_{p,q} \omega_p \omega_q \rho_{p,q}, \tag{22}$$

where we recover (4) and (6) by taking $\omega_p = 1/N$. Moreover, we define the variance of effects as:

$$\hat{\sigma}^2_b = \frac{1}{P-1} \sum_{p=1}^{P} \omega_p (\hat{b}_p - \hat{\mu}_b)^2. \tag{23}$$

To assess the sensitivity of our results to the choice of weighting scheme, we propose below a non-uniform alternative. We posit a baseline path, which we take to be the blue one in Figure 1, which we call path number 1 and to which we assign a score of $s_1 = 1$. All other paths will have a score of $s_p = 0.75^{d(p,1)}$, where $d(p,1)$ is the distance with respect to the initial path, i.e., the number of choices that differ between path p and path 1. Because there are eight possible choices (the eight steps in Figure 1), this means that the maximum distance is equal to eight. In turn, this implies that the minimum score is equal to $s_{\min} = 0.75^8 \approx 0.1$. Thus, some paths, including the orange one in Figure 1, will have a score ten times smaller than the baseline path. The weight of each path is then:

$$\omega_p = \frac{s_p}{\sum_{p=1}^{P} s_p}. \tag{24}$$

In Figure 11 below, we reproduce the analysis from Figure 9 but with the weighted averages and variances defined in Equations (21)-(23) with weights equal to (24). We are only interested in the situations with path-specific resampling. In this case, because of the symmetry in correlations, the term in (22) will not differ much from the baseline situation, and it is (23) that will drive the changes in variability.

From afar, the confidence intervals are similar to those of Figure 9. The ranges of the intervals are roughly unchanged, hence the weighting primarily affects the mean. The most important shift is perhaps that of *IndMom*, with an interval now close to including zero. Nevertheless, the same four factors are found to be sensitive to methodological changes. This indicates that our results are mildly sensitive to the choice of weights. Yet, the latter should be chosen carefully, reflecting an informed judgment on the relative representativeness of the paths.

In Figure 12, we also show the effect of weighting on variance decomposition. There is one notable difference, compared to uniform weights (Figure 10): the ordering is not exactly the same. For instance, *realestate* now ranks fourth, whereas it was second in the original figure. Hence, weighting does mildly alter standard errors as well. Nevertheless, standard errors are roughly comparable to those of Figure 10, highlighting that, in this example, non-uniform weights do not shift the relative importance of SE vs. NSE.

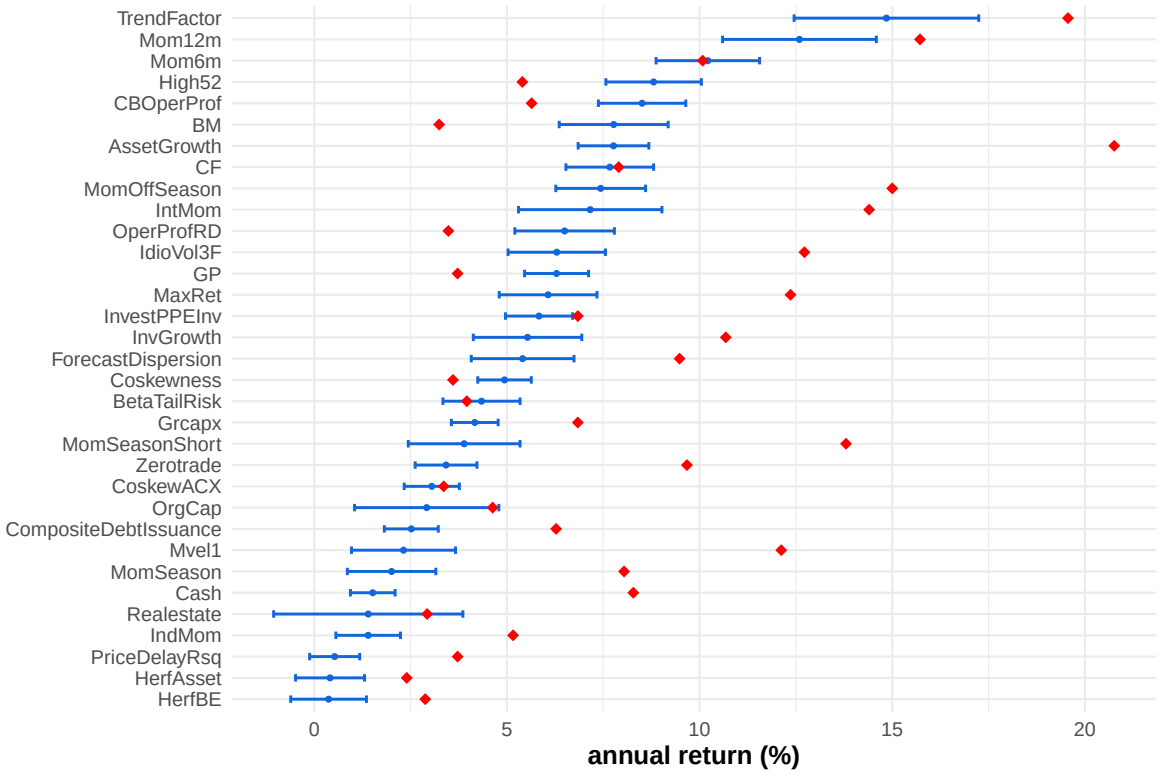

Figure 11: **Resilient anomalies with non-uniform weights**. We replicate the analysis in Figure 9 but with non-uniform weights. The baseline path is the blue one in Figure 1 and each path has a weight proportional to 0.75^d, where d is the distance to the baseline path defined in Equation (24). Averages and variances of outcomes are computed according to Equations (21) and (22).

6.2 Differential sampling

In our paper, the baseline sampling strategy involves randomly selecting 40% of the original data. We adopted this approach for its simplicity, and our various experiments demonstrate that it performs well. Notably, when the sampling is path-specific, the two rightmost terms of Equation (12) (restated below) remain moderate, resulting in the tight confidence intervals shown in Figure 9:

$$\Delta_\alpha = \frac{2\sqrt{3}\sigma_{\hat{\mu}_b}}{\sqrt{\alpha}} \leqslant \frac{2\sqrt{3}\sigma_*}{\sqrt{\alpha}}\sqrt{\frac{2\sqrt{3}\kappa}{\sqrt{\alpha N}} + \sum_{p,q}\frac{\hat{\rho}_{p,q}}{P^2}}.$$

Admittedly, alternative sampling methods could also be employed. For instance, a bootstrap strategy can be implemented to generate outcomes that are independent by design. Unlike in our baseline, there is no need to estimate correlations across paths, since they are known ex ante to be zero, as samples are independent by construction. However, a limitation of bootstrapping methods is that they typically rely on assumptions about the data-generating process.

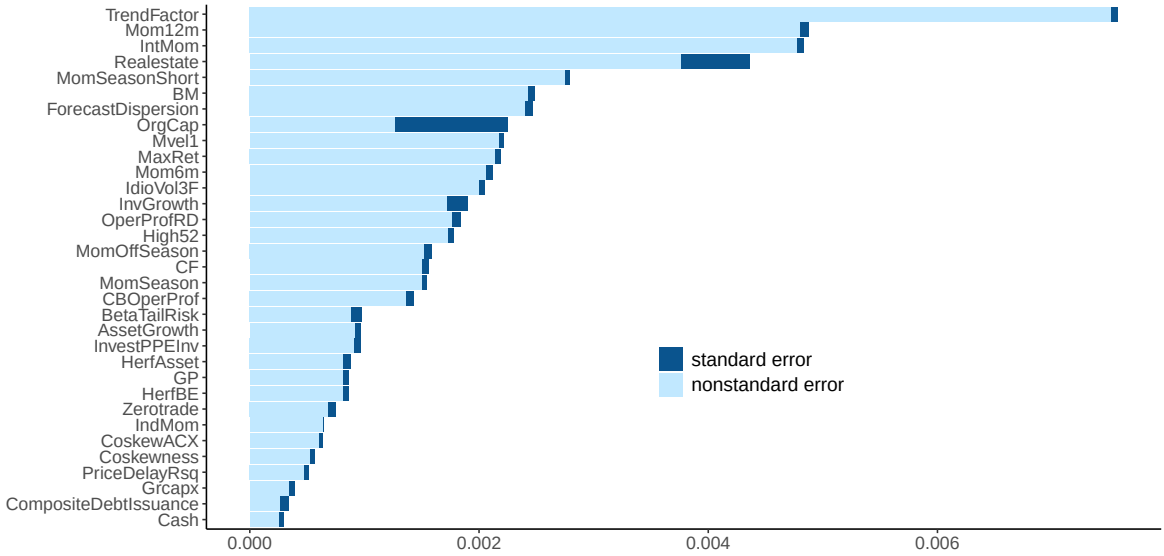

Figure 12: **Variance decomposition under non-uniform weights**. We display the decomposition of the variance in average returns proposed in Equation (19) and formalized in Appendix E, but with weights given in Equation (24). The names of characteristics are those of Chen and Zimmermann (2022a).

When the outcomes are independent by design, the expression for the confidence interval simplifies considerably. Indeed, in this case, there is no need to estimate the correlations between paths, $\rho_{p,q}$, as we know they are null. Relative to Equation (12), two simplifications follow: (1) we can set $N = \infty$ as there is no estimation error, and (2) the average of the correlation terms is $\sum_{p,q} \rho_{p,q}/P^2 = 1/P$ since only the P diagonal terms contribute. Consequently, Equation (12) reduces to:

$$\Delta_\alpha \leqslant \frac{2\sqrt{3}\sigma_*}{\sqrt{\alpha P}}. \tag{25}$$

One popular bootstrap method that allows us to generate independent outcomes is the wild bootstrap (see Davidson and Flachaire (2008)). Before implementing it, we recall that average returns from sorted long-short portfolios can be viewed as estimates from linear regressions without intercepts (see page 326 of Fama (1976) and the Appendix of Freyberger et al. (2020)). We can write $r = xb + e$, where r is the vector of returns and x the sorting variable (in panel format, i.e., across several dates and firms). We can estimate $\hat{b}$ and the corresponding residuals $\hat{e}$, and then generate new returns as follows $r^* = x\hat{b} + u \times \hat{e}$, with u being a vector of iid $\mathcal{N}(0,1)$ variates. The noise term $u \times \hat{e}$ has the same mean and variance as the original errors $\hat{e}$.[12]

[12]We assume iid variables primarily for reasons of simplicity. Since returns are sampled monthly, temporal correlation tends to be weak (Campbell et al. (1997)). Cross-sectional dependence, however, is a more significant concern. The dataset is highly unbalanced, with numerous firms entering and exiting over time, which complicates the estimation of cross-sectional correlations. To address this challenge, we adopt the sim-

We plot in Figure 13 the confidence intervals at the 95% level for all anomalies under wild bootstrap sampling. Both the number and identity of the persistent anomalies are quite similar compared to our baseline results displayed in Figure 9. Indeed, the *PriceDelayRsq*, *HerfAsset*, and *HerfBE* are classified as persistent with both sampling methods. In contrast, the *Realestate* anomaly is only persistent under wild bootstrap but not under our baseline subsampling strategy. This shows that the estimation error from path-specific sampling only had a marginal impact on our conclusions.

Another takeaway from Figure 13 is that interval widths are more homogeneous, compared to those of Figures 9 and 11. This is because the sampling noise is more pronounced with the bootstrap, and it attenuates the differences across anomalies, which, in the previous plots, originated mostly from the paths. Nevertheless, the average effects (i.e., the centers of the intervals) remain invariant across the sampling schemes.

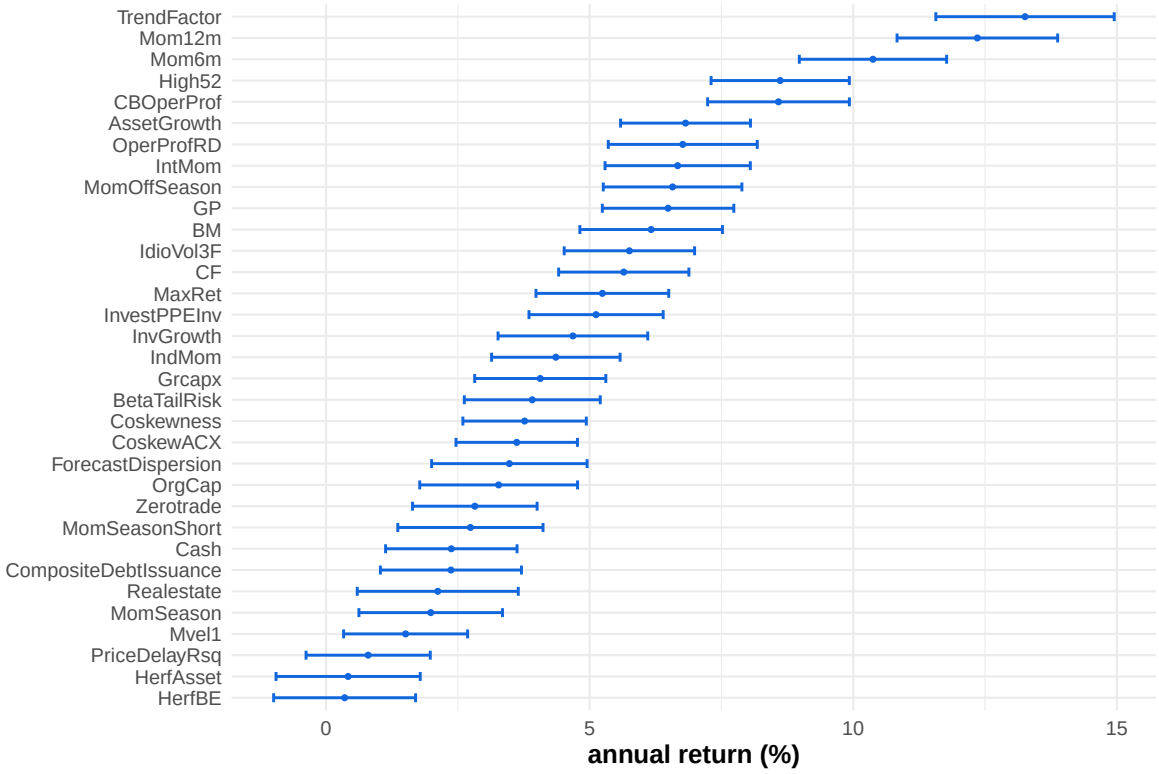

Figure 13: **Resilient anomalies under wild bootstrap sampling**. We plot, for $\alpha = 0.05$, the confidence intervals (7) of the mean of long-short returns under path-specific resampling without estimation error. The width of confidence interval is in this case given by Equation (25). By construction, the sample means lie in the middle of the intervals .

plifying assumption that error terms are stationary and mutually independent, both over time and across firms.

6.3 Enhanced robustness checks

The framework introduced in this paper can also be employed to conduct enhanced robustness checks in any empirical finance analysis. To illustrate this, consider a baseline specification that examines the effect of a variable X_t on an outcome variable Y_{t+1}, while controlling for a set of contemporaneous covariates. This baseline model may be theoretically motivated, derived from an identification strategy, based on prior research, or constructed in a more ad hoc manner. The coefficient of interest, denoted β_0, is associated with the variable X_t, and the main conclusions typically concern its sign, magnitude, and statistical significance.

As shown in Figure 1, this baseline specification corresponds to a single methodological path among many. In the final section of most empirical studies, researchers conduct robustness checks by introducing deviations from the baseline one at a time (e.g., changing the sample, estimation method, set of control variables) while holding all other modeling choices constant. Each such deviation alters one of the J methodological steps and corresponds to a specific alternative path among the P possible paths. This "what-if" analysis yields an alternative estimate, $\hat{\beta}_a$, which is then compared with the baseline result $\hat{\beta}_0$ to assess sensitivity. However, these checks are typically *ceteris paribus* in nature: they test the sensitivity of the result to one decision at a time, keeping all others fixed. A more general approach is to consider vectors of shocks that affect multiple methodological choices simultaneously. However, each such combination also defines a single methodological path.

The framework proposed in this paper allows researchers to go beyond a limited set of ad hoc alternatives by systematically evaluating results across the full space of methodological paths. Rather than focusing on a few hand-picked alternatives, this approach yields a distribution of estimates, providing richer information about the stability and reliability of the baseline result. For example, researchers can report the proportion of specifications that yield positive and statistically significant estimates. Moreover, by applying the *path-specific* sampling strategy described in Section 4.2, one can construct a 95% confidence interval for this proportion, offering a formal measure of inference over robustness.

7 Conclusion

Menkveld et al. (2024) is a truly *path-breaking* paper. Not only in the conventional sense of being highly influential, but also literally, as it *breaks the path* of a single empirical approach and maps out methodological alternatives. We build on their approach to derive a rigorous framework for inference on the average return of a portfolio. Specifically, we derive a canonical decomposition of the variance of the average return. This formula allows us to clearly identify the drivers of the width of the confidence interval around this mean effect. We find that three components matter.

The first component is the sum of the correlation terms across all path outcomes. As this sum shrinks, so does the range of the confidence interval. The second component is the standard error (SE), which quantifies the variation of the effect that are due to sampling

noise. Finally, the third component is the nonstandard error (NSE) that results from the uncertainty generated by methodological choices. Unlike other multi-design studies, we derive both SE and NSE within a unified framework, enabling a meaningful comparison between them.

Empirically, we illustrate these concepts in the context of asset pricing anomalies. Our results show that keeping the data fixed while spanning the paths is detrimental to accuracy because the correlations across paths are strongly skewed to the right. Resorting to resampling allows us to shrink the width of confidence intervals by a factor of three at least. Moreover, assuming a symmetric distribution of correlations further curtails the range of these intervals threefold. These findings underline how crucial resampling can be in multi-design studies. In our study, NSE are much larger than their standard counterparts for most anomalies. Implementing our full methodology allows us to identify 29 *persistent* factors, that are robust to multiple methodological variations. For all of them, the 95% confidence interval for the average return of the long-short portfolio does not include zero.

Overall, we find that the NSE component is the primary determinant of the width of confidence intervals for multi-path average effects. This highlights the need for researchers to more systematically account for uncertainty stemming from methodological choices in their scientific protocols. This paper offers a practical, operational framework to do so. While the improvements we mention are contingent on our dataset, it is likely that similar gains could be obtained in empirical corporate finance (Mitton, 2022), as well as in other scientific areas.

References

Abadie, A., Athey, S., Imbens, G. W., and Wooldridge, J. M. (2020). Sampling-based versus design-based uncertainty in regression analysis. *Econometrica*, 88(1):265–296.

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

Barras, L., Scaillet, O., and Wermers, R. (2010). False discoveries in mutual fund performance: Measuring luck in estimated alphas. *Journal of Finance*, 65(1):179–216.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300.

Beyer, V. and Bauckloh, T. (2024). Non-standard errors in carbon premia. *SSRN Working Paper*, 4901081.

Blitzstein, J. K. and Hwang, J. (2019). *Introduction to probability*. Chapman and Hall/CRC.

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., et al. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, 582:84–88.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 26:123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

Breznau, N., Rinke, E., Wuttke, A., Adem, M., Adriaans, J., Akdeniz, E., Alvarez-Benjumea, A., Andersen, H., Auer, D., Azevedo, F., et al. (2024). The reliability of replications: A study in computational reproductions. *SocArXiv Working Paper*.

Breznau, N., Rinke, E. M., Wuttke, A., Nguyen, H. H., Adem, M., Adriaans, J., Alvarez-Benjumea, A., Andersen, H. K., Auer, D., Azevedo, F., 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., Lé, M., Sangnier, M., and Zylberberg, Y. (2016). Star wars: The empirics strike back. *American Economic Journal: Applied Economics*, 8(1):1–32.

Browne, M. and Shapiro, A. (1986). The asymptotic covariance matrix of sample correlation coefficients under general conditions. *Linear Algebra and its Applications*, 82:169–176.

Bryzgalova, S., Huang, J., and Julliard, C. (2023). Bayesian solutions for the factor zoo: We just ran two quadrillion models. *Journal of Finance*, 78(1):487–557.

Bühlmann, P. (2011). Bagging, boosting and ensemble methods. In *Handbook of computational statistics: Concepts and methods*, pages 985–1022. Springer.

Cakici, N., Fieberg, C., Neszveda, G., Piljak, V., and Zaremba, A. (2025a). Lost in the multiverse: Methodological uncertainty in studying global equity returns. *SSRN Working Paper*, 5181455.

Cakici, N., Fieberg, C., Neumaier, T., Poddig, T., and Zaremba, A. (2025b). The devil in the details: How sensitive are "pockets of predictability" to methodological choices? *Critical Finance Review*, Forthcoming.

Campbell, J. Y., Lo, A. W., MacKinlay, A. C., and Whitelaw, R. F. (1997). *The econometrics of financial markets*. Princeton University Press.

Chen, A. Y. (2021). The limits of p-hacking: Some thought experiments. *Journal of Finance*, 76(5):2447–2480.

Chen, A. Y. and McCoy, J. (2024). Missing values handling for machine learning portfolios. *Journal of Financial Economics*, 155:103815.

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

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

Chen, M., Hanauer, M., and Kalsbach, T. (2025). Design choices, machine learning, and the cross-section of stock returns. *SSRN Working Paper*, 5031755.

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

Cirulli, A., Traut, J., De Nard, G., and Walker, P. S. (2025). Low risk, high variability: Practical guide for portfolio construction. *SSRN Working Paper*, 5105457.

Cohn, J. B., Liu, Z., and Wardlaw, M. I. (2023). Count (and count-like) data in finance. *Journal of Financial Economics*, 146(2):529–551.

Davidson, R. and Flachaire, E. (2008). The wild bootstrap, tamed at last. *Journal of Econometrics*, 146(1):162–169.

Dick-Nielsen, J., Feldhutter, P., Pedersen, L. H., and Stolborg, C. (2023). Corporate bond factors: Replication failures and a new framework. *SSRN Working Paper*, 4586652.

Dickerson, A., Mueller, P., and Robotti, C. (2023). Priced risk in corporate bonds. *Journal of Financial Economics*, 150(2):103707.

Duroux, R. and Scornet, E. (2018). Impact of subsampling and tree depth on random forests. *ESAIM: Probability and Statistics*, 22:96–128.

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

Fama, E. (1976). *Foundations of Finance*. Basic Books, New York.

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

Fama, E. F. and French, K. R. (1996). Multifactor explanations of asset pricing anomalies. *Journal of Finance*, 51(1):55–84.

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

Fieberg, C., Günther, S., Poddig, T., and Zaremba, A. (2024). Non-standard errors in the cryptocurrency world. *International Review of Financial Analysis*, 92:103106.

Freyberger, J., Neuhierl, A., and Weber, M. (2020). Dissecting characteristics nonparametrically. *Review of Financial Studies*, 33(5):2326–2377.

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

Gnambs, T. (2023). A brief note on the standard error of the Pearson correlation. *Collabra: Psychology*, 9(1):87615.

Gould, E., Fraser, H. S., Parker, T. H., Nakagawa, S., Griffith, S. C., Vesk, P. A., Fidler, F., Abbey-Lee, R. N., Abbott, J. K., Aguirre, L. A., et al. (2023). Same data, different analysts: variation in effect sizes due to analytical decisions in ecology and evolutionary biology. *BMC Biology*, 23(35).

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

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

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

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

Heath, D., Ringgenberg, M. C., Samadi, M., and Werner, I. M. (2023). Reusing natural experiments. *Journal of Finance*, 78(4):2329–2364.

Hellum, O., Jensen, T. I., Kelly, B. T., and Pedersen, L. H. (2025). The power of the common task framework. *SSRN Working Paper*, 5242901.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30.

Holzmeister, F., Johannesson, M., Böhm, R., Dreber, A., Huber, J., and Kirchler, M. (2024). Heterogeneity in effect size estimates. *Proceedings of the National Academy of Sciences*, 121(32):e2403490121.

Horowitz, J. L. (2019). Bootstrap methods in econometrics. *Annual Review of Economics*, 11(1):193–224.

Hou, K., Xue, C., and Zhang, L. (2015). Digesting anomalies: An investment approach. *Review of Financial Studies*, 28(3):650–705.

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

Huber, C., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Weitzel, U., Abellán, M., Adayeva, X., Ay, F. C., Barron, K., 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.

Huntington-Klein, N., Arenas, A., Beam, E., Bertoni, M., Bloem, J. R., Burli, P., Chen, N., Greico, P., Godwin, E., Todd, P., Saavedra, M., and Stopnitzky, Y. (2021). The influence of hidden researcher decisions in applied microeconomics. *Economic Inquiry*, 59:944–960.

Huntington-Klein, N., Portner, C. C., McCarthy, I., et al. (2025). The sources of researcher variation in economics. *NBER Working Paper*, 33729.

Ion, R. A., Klaassen, C. A., and Heuvel, E. R. v. d. (2023). Sharp inequalities of Bienaymé–Chebyshev and Gauß type for possibly asymmetric intervals around the mean. *TEST*, pages 1–36.

Jensen, T. I., Kelly, B. T., and Pedersen, L. H. (2023). Is there a replication crisis in finance? *Journal of Finance*, 78(5):2465–2518.

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

Kelly, B. T. and Malamud, S. (2025). Understanding the virtue of complexity. *SSRN Working Paper*, 5346842.

McLean, R. D. and Pontiff, J. (2016). Does academic publication destroy stock return predictability? *Journal of Finance*, 71(1):5–32.

Menkveld, A., Dreber, A., Holzmeister, F., Johannesson, M., Huber, J., Kirchler, M., Neususs, S., Razen, M., Weitzel, U., et al. (2024). Nonstandard errors. *Journal of Finance*, 79(3):2339–2390.

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

Nagel, S. (2019). Replication papers in the JF: An update. *Journal of Finance (Editorial)*.

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

Pérignon, C., Akmansoy, O., Hurlin, C., Dreber, A., Holzmeister, F., Huber, J., Johannesson, M., Kirchler, M., Menkveld, A. J., Razen, M., et al. (2024). Computational reproducibility in finance: Evidence from 1,000 tests. *Review of Financial Studies*, 37(11):3558–3593.

Roberts, M. R. and Whited, T. M. (2013). Endogeneity in empirical corporate finance. volume 2 of *Handbook of the Economics of Finance*, pages 493–572. Elsevier.

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., et al. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, 1(3):337–356.

Soebhag, A., van Vliet, B., and Verwijmeren, P. (2024). Non-standard errors in asset pricing: Mind your sorts. *Journal of Empirical Finance*, 78:101517.

Timmermann, A. (2006). Forecast combinations. *Handbook of economic forecasting*, 1:135–196.

Walter, D., Weber, R., and Weiss, P. (2024). Methodological uncertainty in portfolio sorts. *SSRN Working Paper*, 4164117.

Wang, X., Hyndman, R. J., Li, F., and Kang, Y. (2023). Forecast combinations: An over 50-year review. *International Journal of Forecasting*, 39(4):1518–1547.

Zaman, F. and Hirose, H. (2009). Effect of subsampling rate on subbagging and related ensembles of stable classifiers. In *Pattern Recognition and Machine Intelligence: Third International Conference*, pages 44–49. Springer.

Zhou, Z.-H. (2025). *Ensemble methods: Foundations and Algorithms (Second Edition)*. CRC Press.

A Data

Variable	Description	Return (%)	Authors	Journal	Year
AssetGrowth	Asset growth	1.73	Cooper, Gulen and Schill	JF	2008
BetaTailRisk	Tail risk beta	0.33	Kelly and Jiang	RFS	2014
BM	Book to market, original	0.27	Stattman	Other	1980
Cash	Cash to assets	0.69	Palazzo	JFE	2012
CBOperProf	Cash-based operating profitability	0.47	Ball et al.	JFE	2016
CF	Cash flow to market	0.66	Lakonishok, Shleifer, Vishny	JF	1994
CompositeDebtIssuance	Composite debt issuance	0.52	Lyandres, Sun and Zhang	RFS	2008
CoskewACX	Coskewness using daily returns	0.28	Ang, Chen and Xing	RFS	2006
Coskewness	Coskewness	0.30	Harvey and Siddique	JF	2000
ForecastDispersion	EPS Forecast Dispersion	0.79	Diether, Malloy and Scherbina	JF	2002
GP	gross profits / total assets	0.31	Novy-Marx	JFE	2013
Grcapx	Change in capex (two years)	0.57	Anderson and Garcia-Feijoo	JF	2006
HerfAsset	Industry concentration (assets)	0.20	Hou and Robinson	JF	2006
HerfBE	Industry concentration (equity)	0.24	Hou and Robinson	JF	2006
High52	52 week high	0.45	George and Hwang	JF	2004
IdioVol3F	Idiosyncratic risk (3 factors)	1.06	Ang et al.	JF	2006
IndMom	Industry Momentum	0.43	Grinblatt and Moskowitz	JF	1999
IntMom	Intermediate Momentum	1.20	Novy-Marx	JFE	2012
InvestPPEInv	change in ppe and inv/assets	0.57	Lyandres, Sun and Zhang	RFS	2008
InvGrowth	Inventory Growth	0.89	Belo and Lin	RFS	2012
MaxRet	Maximum return over month	1.03	Bali, Cakici, and Whitelaw	JFE	2011
Mom12m	Momentum (12 month)	1.31	Jegadeesh and Titman	JF	1993
Mom6m	Momentum (6 month)	0.84	Jegadeesh and Titman	JF	1993
MomOffSeason	Off season long-term reversal	1.25	Heston and Sadka	JFE	2008
MomSeason	Return seasonality years 2 to 5	0.67	Heston and Sadka	JFE	2008
MomSeasonShort	Return seasonality last year	1.15	Heston and Sadka	JFE	2008
OperProfRD	Operating profitability R&D adjusted	0.29	Ball et al.	JFE	2016
OrgCap	Organizational capital	0.39	Eisfeldt and Papanikolaou	JF	2013
PriceDelayRsq	Price delay r square	0.31	Hou and Moskowitz	RFS	2005
Realestate	Real estate holdings	0.24	Tuzel	RFS	2010
Mvel1	Size	1.01	Banz	JFE	1981
TrendFactor	Trend Factor	1.63	Han, Zhou, Zhu	JFE	2016
Zerotrade	Days with zero trades	0.81	Liu	JFE	2006

Table A.1: **Information on anomalies**. We provide descriptions and sources for the 33 asset pricing anomalies examined in this paper. Journal abbreviations are as follows: *JF = Journal of Finance*; *JFE = Journal of Financial Economics*; *RFS = Review of Financial Studies*. The data, including the original average monthly returns reported in the respective studies (see column headed Return (%)), are sourced from the Open Source Asset Pricing project.

B Details on portfolio construction

This appendix outlines the steps involved in transforming the raw data into average returns for the long-short portfolios constructed around various market anomalies. As depicted in Figure 1, we consider alternative forks for seven decisions:

1. **Financials**. Financial firms (e.g., banks or insurance companies) have different business models and capital structures, compared to other firms. In particular, they are highly leveraged. Therefore, some authors, including Fama and French (1992), prefer to exclude these companies from their analysis. Others choose to include them. Exclusion is usually performed via the Standard Industrial Classification (SIC) classification and all firms with SIC codes between 6000 and 6999 can be withdrawn from the data before continuing the process.
2. **Size filter**. Asset pricing anomalies are only interesting to investors if they imply realistic strategies. One major issue is that of liquidity, i.e., the ability to buy and sell stocks easily without being exposed to large bid-ask spreads. For this reason, it is customary to exclude the least liquid stocks from the analysis. This is done by filtering either according to price (removing penny stocks) or market capitalization (withdrawing the smallest firms). In this paper, we adopt the latter approach by excluding either the bottom 5% or 10% of firms with the smallest market capitalizations each month.
3. **Imputation**. Missing data are ubiquitous in financial datasets. However, their prevalence is also an obstacle because they often imply to discard many observations, thereby curtailing the amount of information to be analyzed. To avoid this, it is customary to replace the missing points by an estimate of the best guess for these points. Following Gu et al. (2020) and Chen and McCoy (2024)), we use cross-sectional averages or medians, which are computed on a monthly basis.
4. **Quantile threshold**. When exploiting anomalies, sorting the is standard procedure. Each month, the stocks are ranked based on a specific firm characteristic, such as firm size, book-to-market, etc. Then, the top X% and bottom X% of stocks, according to this characteristic are grouped into two separate portfolios, the long portfolio and the short portfolio. In the academic literature, X is typically set to 10 or 20, meaning that each leg (long or short) comprises exactly one-tenth or one-fifth of the entire stock universe at any given date. In this paper, we adhere to these two commonly used values.
5. **Sample period**. Full-sample averages can be misleading. For instance, suppose an anomaly yields an average return of +1% per month over the entire sample, but delivers +3% in the first half and -1% in the second half. In such a case, the anomaly is unlikely to be persistent, and investors may be hesitant to act on it. To address this, researchers commonly perform subsample analyses as robustness checks. We adopt this approach by considering six distinct periods, defined by three starting points (the beginning, the first third, and the second third of the sample) and three corresponding endpoints (the first third, the second third, and the end of the sample). Prior to analysis, we restrict the data to include only the observations within the selected period.

6. **Weighting scheme**. After selecting the stocks for inclusion in the long and short portfolios, the next step is to assign portfolio weights. These weights are typically either equal-weighted—giving each stock the same importance—or value-weighted, i.e., assigning weights in proportion to market capitalization. Equal-weighting enhances diversification by treating all stocks equally, while value-weighting mirrors the structure of major equity indices, where larger firms have greater influence due to their larger market share. In this paper, we employ both approaches.
7. **Holding period**. After constructing the portfolios, the holding period—i.e., the length of time before rebalancing—is flexible. Typically, portfolios are rebalanced at regular intervals measured in months: commonly every month, every three months (quarterly), or every twelve months (annually). In this paper, we adopt monthly and quarterly rebalancing frequencies for our analysis.

C Inference on the mean

Let us recall Equation (10):

$$\mathbb{P}\left[|\hat{\mu}_b - \mu_b| \leqslant \frac{2\sigma_{\hat{\mu}_b}}{3\sqrt{\alpha}}\right] \geqslant 1-\alpha,$$

which relies on $\sigma_{\hat{\mu}_b}$, and this value depends mostly on the information on the correlations $\rho_{p,q}$. Gnambs (2023) suggests that a reasonable choice for the variance of the estimator of the correlation is $(1-\rho_{p,q}^2)/\sqrt{N-3}$, where N is the number of samples that are generated to compute the correlations. In any case, it is evident that there exists a N^* such that for $N \geqslant N^*$, it holds that:

$$\sigma^2_{\hat{\rho}_{p,q,N}} \leqslant N^{-1}, \tag{26}$$

where the N index underlines the fact that estimations are performed on samples of size N. We will henceforth assume that this inequality holds. For $N \geqslant 100$, the simulation studies in Gnambs (2023) suggest that the error on the standard error of correlations is marginal. Moreover, by Lemma 1 below, we have, for large N:

$$\sigma^2_{\bar{\rho}} = \mathbb{V}\left[\sum_{p,q}\frac{\hat{\rho}_{p,q,N}}{P^2}\right] \to \frac{\kappa^2}{N},$$

where the constant κ^2 only depends on the correlations $\rho_{p,q}$. The Bienaymé-Chebyschev inequality, applied to the estimated correlation implies, again for large enough N:

$$\mathbb{P}\left[P^{-2}\left|\sum_{p,q}\rho_{p,q} - \hat{\rho}_{p,q}\right| \leqslant v\right] \geqslant 1 - \left(\frac{2\sigma_{\bar{\rho}}}{3v}\right)^2 \geqslant 1 - \frac{4\kappa^2}{9v^2N}.$$

For N large, it holds that:

$$\mathbb{P}\left[P^{-2}\left|\sum_{p,q}\rho_{p,q} - \hat{\rho}_{p,q}\right| \leqslant \frac{2\kappa}{3\sqrt{\alpha N}}\right] \leqslant 1-\alpha, \tag{27}$$

and, with probability $1-\alpha$ at least,

$$\sigma^2_{\hat{\mu}_b} \leqslant \sigma_b^2\left(\frac{2\kappa}{3\sqrt{\alpha N}} + \frac{1}{P^2}\sum_{p,q}\hat{\rho}_{p,q}\right) \leqslant \sigma_*^2\left(\frac{2\kappa}{3\sqrt{\alpha N}} + \sum_{p,q}\frac{\hat{\rho}_{p,q}}{P^2}\right),$$

where we have written σ_*^2 for a known upper bound for σ_b^2 (see Appendix D). Plugging this into Equation (10) yields:

$$\mathbb{P}\left[|\hat{\mu}_b - \mu_b| \leqslant \frac{2\sigma_*}{3\sqrt{\alpha}}\sqrt{\frac{2\kappa}{3\sqrt{\alpha N}} + \sum_{p,q}\frac{\hat{\rho}_{p,q}}{P^2}}\right] \geqslant 1-\alpha, \tag{28}$$

which is the sought interval.

The final step pertains to the coverage of the three confidence bounds: on the correlations, on the sample mean, and on σ_* below. The *adverse* (complement) sets in the probabilities of Equations (27), (28), and (33) are:

$$\begin{aligned}
\mathcal{A}_\rho &= P^{-2}\left|\sum_{p,q}\rho_{p,q}-\hat{\rho}_{p,q}\right| > \frac{2\kappa}{3\sqrt{\alpha N}} \\
\mathcal{A}_\mu &= |\hat{\mu}_b-\mu_b| > \frac{2\sigma_*}{3\sqrt{\alpha}}\sqrt{\frac{2\kappa}{3\sqrt{\alpha N}}+\sum_{p,q}\frac{\hat{\rho}_{p,q}}{P^2}} \\
\mathcal{A}_\sigma &= |\sigma_b^2-\hat{\sigma}_b^2| > \frac{\hat{\sigma}_b^2}{1-\epsilon}.
\end{aligned}$$

We want to avoid the union of these three sets (one, or the other, or the third), knowing that the probability of each is smaller than α. Hence, $\mathbb{P}[\mathcal{A}_\mu \cup \mathcal{A}_\sigma \cup \mathcal{A}_\rho] \leqslant 3\alpha$. This implies that we can conservatively replace α in our overarching inequality by $\alpha/3$, so that the global probability is indeed smaller than α. This leads to the adjusted interval:

$$\mathbb{P}\left[|\hat{\mu}_b-\mu_b| \leqslant \frac{2\sqrt{3}\sigma_*}{\sqrt{\alpha}}\sqrt{\frac{2\sqrt{3}\kappa}{\sqrt{\alpha N}}+\sum_{p,q}\frac{\hat{\rho}_{p,q}}{P^2}}\right] \geqslant 1-\alpha, \tag{29}$$

which is the one (Equation (12)) we use throughout the paper.

Our line of reasoning relies on the following technical result.

Lemma 1. *The covariance matrix of correlation coefficients is such that, asymptotically for large* N,

$$\mathbb{V}\left[\sum_{p,q}\frac{\hat{\rho}_{p,q,N}}{P^2}\right] \overset{N\to\infty}{\to} \frac{f(\mathbf{\Sigma})}{N} := \frac{\kappa^2}{N}, \tag{30}$$

where $f(\mathbf{\Sigma})$ *is simply a (scalar) function of correlations specified in the proof below.*

We prove the lemma below. It is also easy to test it numerically with simulations (which are available upon request). First, Corollary 2 from Browne and Shapiro (1986) states that the approximation, for a large sample size N, of the covariance of sample correlations is:

$$\begin{aligned}
\kappa_{i,j,h,k} &= N\mathbb{E}[(\hat{\rho}_{i,j,N}-\rho_{i,j})(\hat{\rho}_{h,k,N}-\rho_{h,k})] \\
&\to \rho_{i,k}\rho_{j,h}+\rho_{i,h}\rho_{j,k}+\frac{1}{2}\sum_{l,m}(\delta_{i,l}+\delta_{j,l})\rho_{i,j}\rho_{l,m}^2\rho_{h,k}(\delta_{h,m}+\delta_{k,m}) \\
&\quad -\sum_{l=1}^{P}\rho_{i,l}\rho_{j,l}(\delta_{k,l}+\delta_{h,l})\rho_{k,h}-\sum_{l=1}^{P}\rho_{k,l}\rho_{h,l}(\delta_{i,l}+\delta_{j,l})\rho_{i,j},
\end{aligned}$$

where $\delta_{i,j} = 1_{\{i=j\}}$ is the Kronecker delta. Hence,

$$\begin{aligned}\sum_{i,j,k,h} \kappa_{i,j,h,k} \to & \sum_{i,j,h,k} \rho_{i,k}\rho_{j,h} + \rho_{i,h}\rho_{j,k} \\ & + \frac{1}{2}\sum_{i,j,k,h}\sum_{l,m}(\delta_{i,l}+\delta_{j,l})\rho_{i,j}\rho_{l,m}^2\rho_{h,k}(\delta_{h,m}+\delta_{k,m}) \\ & - \sum_{i,j,k,h}\sum_{l=1}^{P}\rho_{i,l}\rho_{j,l}(\delta_{k,l}+\delta_{h,l})\rho_{k,h} - \sum_{i,j,k,h}\sum_{l=1}^{P}\rho_{k,l}\rho_{h,l}(\delta_{i,l}+\delta_{j,l})\rho_{i,j} \\ = & \, 2\left(\sum_{i,k}\rho_{i,k}\right)\left(\sum_{j,h}\rho_{j,h}\right) + 2\sum_{i,j,k,h}\rho_{i,j}\rho_{i,k}^2\rho_{h,k} - 4\sum_{i,j,k,l}\rho_{i,l}\rho_{j,l}\rho_{k,l} \\ = & \, 2\left(\sum_{i,k}\rho_{i,k}\right)^2 + 2\sum_{i,j,k,h}\rho_{i,j}\rho_{i,k}^2\rho_{h,k} - 4\sum_{i,j,k,l}\rho_{i,l}\rho_{j,l}\rho_{k,l}\end{aligned}$$

Since

$$\begin{aligned}\sigma_{\bar{\rho}}^2 = \mathbb{V}\left[\sum_{p,q}\frac{\hat{\rho}_{p,q,N}}{P^2}\right] = \frac{1}{P^4}\mathbb{E}\left[\sum_{p,q}(\hat{\rho}_{p,q,N}-\rho_{p,q})^2\right] = \frac{1}{P^4}\sum_{p,q,m,n}\mathbb{E}[(\hat{\rho}_{p,q,N}-\rho_{p,q})(\hat{\rho}_{m,n,N}-\rho_{m,n})] \\ \to \frac{1}{NP^4}\left(2\|\mathbf{\Sigma}\|^2 + 2\sum_{i,j,k,h}\rho_{i,j}\rho_{i,k}^2\rho_{h,k} - 4\sum_{i,j,k,l}\rho_{i,l}\rho_{j,l}\rho_{k,l}\right) := \frac{\kappa^2}{N} \qquad (31)\end{aligned}$$

we obtain the sought result. If the correlations are bounded from above by ρ_+, then the above quantity is smaller than $\frac{2\rho_+^2(1+\rho_+^2)}{N}$. In practice, the variance of average correlations is small. For instance, the expression (31) computed for the (estimated) correlations related to Figure 3 lies between 0.088/500 for the factor size and 0.101/500 for asset growth, with $N = 500$ being the number of samples. For the correlations whose distributions are depicted in Figure 5, the magnitude of the variance $\sigma_{\bar{\rho}}^2$ is $10^{-4}/500$, i.e., it is negligible. We could consider the error stemming from estimates (i.e., plugging $\hat{\rho}_{n,m}$ instead of $\rho_{n,m}$), but for $N \geqslant 500$, the adjustment is inconsequential.

D Discussion on σ_*^2

The true variance of b, σ_b^2 is unknown and estimated with $\hat{\sigma}_b^2$ (the sample variance). The latter is computed across paths, and potentially, across samples too. The issue is that we do not know much about the properties of $\hat{\sigma}_b^2$. In particular, we need to quantify its variance in order to be able to characterize the potential error we are making, compared to σ_b^2. Below, we introduce an assumption that is empirically verified upon path-specific sampling and that allows us to derive a bound on σ_b^2 with a confidence level of $1-\alpha$.

Assumption 2. *It holds that:*

1. *the correlations $\rho_{p,q}$ are symmetric around zero; in particular, their sum is null;*
2. $\sum_{p,q,r} \mathbb{E}[(\hat{b}_p - \mu_b)(\hat{b}_q - \mu_b)(\hat{b}_r - \mu_b)] = 0$.

If paths are jointly uncorrelated, then the hypothesis on co-skewness in the second point is straightforward. We then proceed with a result on the variance of the sample variance. It relies on a quantity, ϱ, which is very close to zero according to our estimations upon path-specific resampling. To substantiate this claim, we provide in Figure D.1 the distribution of the correlation between squared effects for the four baseline anomalies in our study.

Now, suppose we have an unbiased estimation σ_b^2 from P paths. Then, via the Bienaymé-Chebyschev inequality:

$$\mathbb{P}\left[|\sigma_b^2 - \hat{\sigma}_b^2| \leqslant \frac{2}{3}\sqrt{\frac{\mathbb{V}[\hat{\sigma}_b^2]}{\alpha}}\right] \geqslant 1-\alpha, \quad \alpha \in (0,1).$$

Under Assumption 2 and Lemma 2 (stated and proven below), and for $\varrho \approx 0$, we thus have:

$$\mathbb{P}\left[|\sigma_b^2 - \hat{\sigma}_b^2| \leqslant \frac{2\sigma_b^2}{3\sqrt{\alpha}}\sqrt{\frac{2}{(P-1)}}\right] \geqslant 1-\alpha, \quad \alpha \in (0,1). \tag{32}$$

The bound in the bracket depends on σ_b, which is unknown, but we can consider the two cases $\sigma_b^2 < \hat{\sigma}_b^2$ and $\sigma_b^2 > \hat{\sigma}_b^2$:

- if $\sigma_b^2 \leqslant \hat{\sigma}_b^2$, then Equation (32) holds when replacing σ_b^2 with $\hat{\sigma}_b^2$.
- if $\sigma_b^2 \geqslant \hat{\sigma}_b^2$, we rewrite the inequality inside the probability of (32) as $\sigma_b^2 - \hat{\sigma}_b^2 \leqslant \epsilon\sigma_b^2$, with $\epsilon = \frac{2}{3\sqrt{\alpha}}\sqrt{\frac{2}{(P-1)}} \in (0,1)$ for P large enough, hence $(1-\epsilon)\sigma_b^2 \leqslant \hat{\sigma}_b^2$, i.e. $\sigma_b^2 \leqslant \frac{\hat{\sigma}_b^2}{1-\epsilon}$.

In the end, we can replace σ_b^2 by $\frac{\hat{\sigma}_b^2}{1-\epsilon}$ in the right-hand side of the inequality within the probability, and obtain:

$$\mathbb{P}\left[|\sigma_b^2 - \hat{\sigma}_b^2| \leqslant \frac{\hat{\sigma}_b^2}{1-\epsilon}\right] \geqslant 1-\alpha, \quad \alpha \in (0,1), \tag{33}$$

and hence, since σ_b^2 and $\hat{\sigma}_b^2$ are very close with probability $1-\alpha$, we can set:

$$\sigma_*^2 = \hat{\sigma}_b^2 \left(1 + \frac{\hat{\sigma}_b^2}{1 - \frac{2}{3\sqrt{\alpha}}\sqrt{\frac{2}{(P-1)}}} \right) \tag{34}$$

as the sought upper bound.

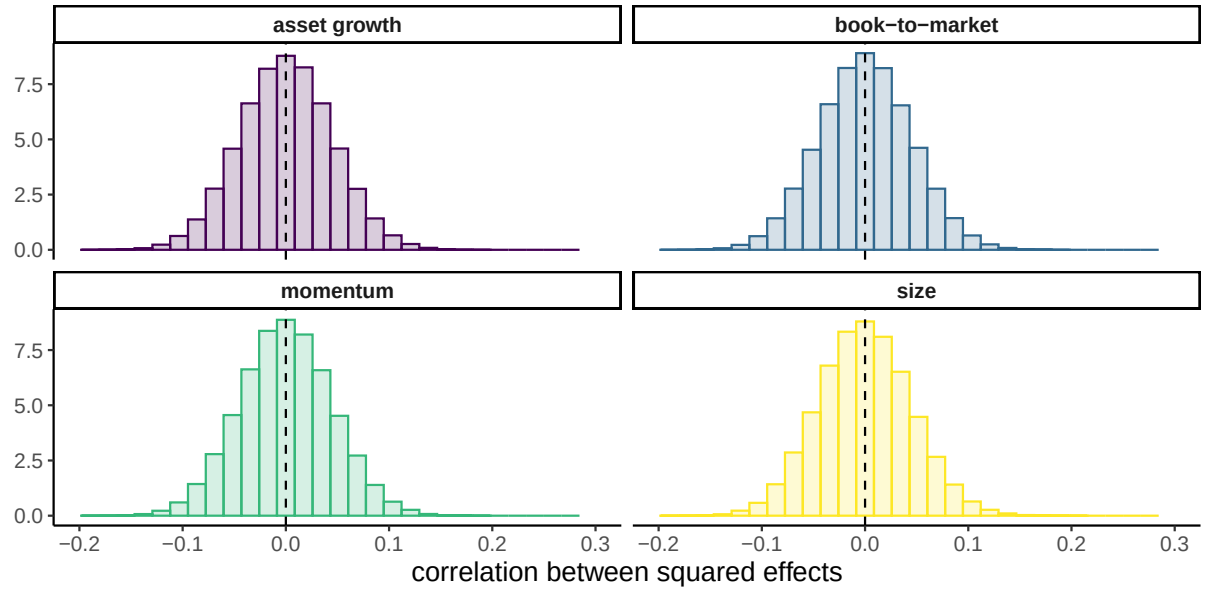

Figure D.1: **Distribution of correlations between *squared* effects**. We show the distribution of the correlations $\mathbb{C}or(\hat{b}_p^2, \hat{b}_q^2)$ for each of the four sorting variables. Correlations are computed on $N = 500$ samples and the samples are generated separately for each path, following our path-specific approach.

Lemma 2. *Under Assumption 2, the variance of the sample variance is* $\mathbb{V}[\hat{\sigma}_b^2] = \sigma_b^4 \left(\frac{2}{P-1} + \varrho\right)$, *where* $\varrho = \frac{1}{(P-1)^2} \sum_{p \neq r} Cor\left(\hat{b}_p^2, \hat{b}_r^2\right)$.

Proof. First, as a side note, we note that:

$$\mathbb{V}[\hat{\mu}_b] = \mathbb{V}\left[\frac{1}{P}\sum_{p=1}^{P} \hat{b}_p\right] = \frac{1}{P^2}\mathbb{E}\left[\sum_{p,q}(\hat{b}_p - \bar{b})^2\right] = \frac{\sigma_b^2}{P^2}\sum_{p,q}\rho_{p,q} = \frac{\sigma_b^2}{P},$$

because the sum of correlation collapses (only the variances remain). Next, we have:

$$\hat{\sigma}_b^2 = \frac{1}{P-1}\sum_{p=1}^{P}\left(\hat{b}_p - \frac{1}{P}\sum_{q=1}^{P}\hat{b}_q\right)^2 = \frac{1}{P-1}\left\{\sum_{p=1}^{P}\hat{b}_p^2 - P\left(\frac{1}{P}\sum_{q=1}^{P}\hat{b}_q\right)^2\right\},$$

and, in addition, it is an unbiased estimator, i.e., $\mathbb{E}[\hat{\sigma}_b^2] = \sigma_b^2$. Indeed,

$$\mathbb{E}[\hat{\sigma}_b^2] = \frac{1}{P-1}\mathbb{E}\left\{\sum_{p=1}^{P}\hat{b}_p^2 - P\left(\frac{1}{P}\sum_{q=1}^{P}\hat{b}_q\right)^2\right\} = \frac{P}{P-1}\left(\sigma_b^2 + \mu_b^2 - (\sigma_b^2/P + \mu_b^2)\right) = \sigma_b^2.$$

Moreover, by Assumption 2, we have:

$$\begin{aligned}\mathbb{V}\left[\sum_{p=1}^{P}\hat{b}_p^2\right] &= \sum_{p=1}^{P}\mathbb{V}\left[\hat{b}_p^2\right] + \underbrace{\sum_{p\neq r}\mathrm{Cov}\left(\hat{b}_p^2, \hat{b}_r^2\right)}_{=\sigma_b^4\varrho(P-1)^2} = \sum_{p=1}^{P}\mathbb{E}\left[\hat{b}_p^4\right] - \mathbb{E}\left[\hat{b}_p^2\right]^2 + \sigma_b^4\varrho(P-1)^2 \\ &= P(3\sigma_b^4 + 6\sigma_b^2\mu_b^2 + \mu_b^4 - (\sigma_b^2 + \mu_b^2)^2) + \sigma_b^2\varrho(P-1)^2 \\ &= P(2\sigma_b^4 + 4\sigma_b^2\mu_b^2) + \sigma_b^4\varrho(P-1)^2\end{aligned}$$

and

$$\begin{aligned}\mathbb{V}\left[\left(\sum_{q=1}^{P}\hat{b}_q\right)^2\right] = \mathbb{V}\left[\sum_{p,q}\hat{b}_p\hat{b}_q\right] &= \mathbb{E}\left[\left(\sum_{p,q}\hat{b}_p\hat{b}_q\right)^2\right] - \left(\mathbb{E}\left[\sum_{p,q}\hat{b}_p\hat{b}_q\right]\right)^2 \\ &= \mathbb{E}\left[\sum_{p,q,r,s}\hat{b}_p\hat{b}_q\hat{b}_r\hat{b}_s\right] - \left(\sum_{p,q}\mathbb{E}\left[\hat{b}_p\hat{b}_q\right]\right)^2 \\ &= 3P^2\sigma_b^4 + P^4\mu_b^4 + 6P^2\sigma_b^2\mu_b^2 - (P\sigma_b^2 + P^2\mu_b^2)^2 \\ &= 2P^2(\sigma_b^4 + 2P\sigma_b^2\mu_b^2),\end{aligned}$$

where the third row above comes from a generalization of Isserlis' theorem to non-central Gaussian laws. If the effects have zero mean, then Isserlis' theorem for $\mathbb{E}[\mu_p\mu_q\mu_r\mu_s]$ and Asssumption 2 imply:

$$\begin{aligned}&\sum_{p,q,r,s}\mathbb{E}[(b_p + \mu_b)(b_q + \mu_b)(b_r + \mu_b)(b_s + \mu_b)] \\ &= \sum_{p,q,r,s}\mathbb{E}[\mu_p\mu_q\mu_r\mu_s] + 4\mu\mathbb{E}[b_pb_qb_r] + 6\mu^2\mathbb{E}[b_pb_q] + 4\mu^2\mathbb{E}[b_p] + \mu^4 \\ &= 3P^2\sigma_b^4 + 6P^2\sigma_b^2\mu_b^2 + P^4\mu_b^4\end{aligned}$$

Finally, we will need the following expression:

$$\begin{aligned}\sum_{p,q,r}\mathrm{Cov}\left(\hat{b}_p^2, \hat{b}_q\hat{b}_r\right) &= \sum_{p,q,r}\mathbb{E}\left[\hat{b}_p^2\hat{b}_q\hat{b}_r\right] - \mathbb{E}[\hat{b}_p^2]\mathbb{E}[\hat{b}_q\hat{b}_r] \\ &= \sigma_b^4(P^2 + 2P) + (P^3 + 5P^2)\mu_b^2\sigma_b^2 + P^3\mu_b^4 - P(\sigma_b^2 + \mu_b^2)(P\sigma_b^2 + P^2\mu_b^2) \\ &= 2P\sigma_b^4 + 4P^2\sigma_b^2\mu_b^2\end{aligned}$$

where we have again resorted to a variation of Isserlis' theorem in the second row. Now,

gathering everything in a bigger picture and aggregating the pieces:

$$
\begin{aligned}
\mathbb{V}[\hat{\sigma}_b^2] &= \mathbb{V}\left[\frac{1}{P-1}\left\{\sum_{p=1}^{P}\hat{b}_p^2 - P\left(\frac{1}{P}\sum_{q=1}^{P}\hat{b}_q\right)^2\right\}\right] = \frac{1}{(P-1)^2}\mathbb{V}\left[\sum_{p=1}^{P}\hat{b}_p^2 - \frac{1}{P}\left(\sum_{q=1}^{P}\hat{b}_q\right)^2\right] \\
&= \frac{1}{(P-1)^2}\left\{\mathbb{V}\left[\sum_{p=1}^{P}\hat{b}_p^2\right] + \frac{1}{P^2}\mathbb{V}\left[\left(\sum_{q=1}^{P}\hat{b}_q\right)^2\right] - \frac{2}{P}\operatorname{Cov}\left(\sum_{p=1}^{P}\hat{b}_p^2, \left(\sum_{q=1}^{P}\hat{b}_q\right)^2\right)\right\} \\
&= \frac{1}{(P-1)^2}\left\{P(2\sigma_b^4 + 4\sigma_b^2\mu_b^2) + \sigma_b^4\varrho(P-1)^2 + 2(\sigma_b^4 + 2P\sigma_b^2\mu_b^2) - \frac{2}{P}\sum_{p,q,r}\operatorname{Cov}\left(\hat{b}_p^2, \hat{b}_q\hat{b}_r\right)\right\} \\
&= \frac{1}{(P-1)^2}(2\sigma_b^4(P+1) + 8P\sigma_b^2\mu_b^2 - 4(\sigma_b^4 + 2P^2\sigma_b^2\mu_b^2)) + \sigma_b^4\varrho \\
&= \frac{2\sigma_b^4}{P-1} + \sigma_b^4\varrho.
\end{aligned}
$$

□

E Computation of SE and NSE

In this section, we provide the sample-based definition of Equations (17) and (18). We present the general formulation with potentially non-uniform weights ω_p that sum to one. Below, $\hat{b}_p(\mathbb{D}_n)$ is related to path p and to subsample n.

$$
\text{SE} = \sqrt{1/2}\sqrt{\sum_{p=1}^{P}\omega_p\hat{\sigma}_p^2 + \frac{1}{N}\sum_{n=1}^{N}(\hat{\mu}_n - \hat{\mu})^2} \tag{35}
$$

$$
\text{NSE} = \sqrt{1/2}\sqrt{\frac{1}{N}\sum_{n=1}^{N}\hat{\sigma}_n^2 + \sum_{p=1}^{P}\omega_p(\hat{\mu}_p - \hat{\mu})^2}, \tag{36}
$$

where

$$
\hat{\mu}_p = \frac{1}{N}\sum_{n=1}^{N}\hat{b}_p(\mathbb{D}_n), \quad \hat{\mu}_n = \sum_{p=1}^{P}\omega_p\hat{b}_p(\mathbb{D}_n), \quad \hat{\mu} = \sum_{p=1}^{P}\omega_p\hat{\mu}_p = \frac{1}{N}\sum_{n=1}^{N}\hat{\mu}_n
$$

and

$$
\hat{\sigma}_n^2 = \sum_{p=1}^{P}\omega_p\left(\hat{b}_p(\mathbb{D}_n) - \hat{\mu}_p\right)^2, \quad \hat{\sigma}_p^2 = \frac{1}{N}\sum_{n=1}^{N}\left(\hat{b}_p(\mathbb{D}_p) - \hat{\mu}_n\right)^2.
$$