Anomaly Persistence and Nonstandard Errors

Guillaume Coqueret* Christophe Pérignon†

May 29, 2025

Abstract

This article presents a framework for robust inference that accounts for the many methodological choices involved in testing asset pricing anomalies. We demonstrate that running multiple paths on the same original dataset inherently results in high correlation across outcomes, which significantly alters inference. In contrast, path-specific resampling greatly reduces outcome correlations and tightens the confidence interval of the estimated average effect. Jointly accounting for across-path and within-path variability allows the variance of the average effect to be decomposed into a standard error, a nonstandard error, and a correlation term. In our empirical analysis, we find that 29 anomalies can be classified as persistent, as their 95% confidence intervals for average returns exclude zero. Our results also indicate that for most anomalies, nonstandard errors dwarf standard errors and are the primary determinants of the width of confidence intervals for multi-path average effects.

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 below), building on emerging practices in other scientific disciplines.[1] Such 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, it provides a more comprehensive characterization of the analyzed phenomenon and serves as a potential remedy for p-hacking in empirical research (Chen, 2021).[2] Second, it quantifies the uncertainty arising from ad hoc methodological choices made by researchers, which Menkveld et al. (2024) coined as *nonstandard errors* (NSE).

Multi-design studies offer valuable opportunities for advancing asset pricing. Indeed, the debate surrounding the robustness of empirical findings and the uncertainty stemming from methodological decisions is particularly intense regarding the so-called *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; Zhang et al., 2025).

Despite growing interest in multi-design studies, rigorous approaches for handling the multitude of resulting estimates remain underdeveloped. In this paper, we propose a framework that enables both methodological and practical contributions to the literature. Methodologically, our main contribution is to demonstrate how to formally test whether the average effect differs from zero. To do so, we decompose of the variance of the mean effect into the correlation terms among the various generated estimates, the standard error (SE) and the NSE of the effect. In contrast with 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. On the practical side, we apply these results to the context of asset pricing anomalies. In particular, we derive and implement statistical tests for identifying *persistent anomalies*—those that are robust to methodological variations.

The various steps of our analysis are the following. We start by showing that overlapping research 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. To address this, we demonstrate that path-specific resampling significantly reduces outcome correlations and symmetrizes their distribution around zero, leading to tighter confidence intervals.

[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) and Breznau et al. (2022, 2024) 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).

To see why this happens, we recall that the width of the confidence intervals around the mean increases with the variance of the estimated mean. We show that the variance of the sample mean of outcomes 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_b^2. Therefore, as the correlations between outcome increase, the precision of the estimates decreases because of the first term.

The second term is also important and we resort to the law of total variance to decompose the variance of the effect into two components: $\sigma_b^2 = \text{SE}^2 + \text{NSE}^2$. Here, SE captures the contribution of sampling effects and NSE reflects the contribution of methodological variation. As they both represent a fraction of the same variance, they are directly comparable. This allows us to propose a canonical decomposition of the variance of the mean effect: $\sigma_{\hat{\mu}_b}^2 = (\text{SE}^2 + \text{NSE}^2) \sum_{p,q} \rho_{p,q} / P^2$.

In an empirical analysis of 33 asset pricing anomalies, we consider eight critical methodological choices (e.g., sample period, holding period, long-short quantiles) for a total of 576 research 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 follow two resampling strategies, both involving 500 new samples: (1) all paths are run on the same new samples (*common resampling* strategy); and (2) new samples are drawn separately for each individual path (*specific resampling* strategy).

We show that the resulting average correlation with common resampling is around 30% whereas it is below 0.25% with specific resampling. From the canonical decomposition of the variance, we directly see the advantage of resorting to specific resampling is to shrink the variance more than 100 times. Consequently, the confidence interval for the mean effect is reduced approximately by a factor of ten. Applying our strategy to the 33 anomalies reveals 29 that can be classified as persistent, with the strongest effects linked to trend-following and momentum strategies.

The 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. Finally, we show that our methodology can accommodate non-uniform weighting schemes across paths reflecting preferences and/or theoretical guidance. To this purpose, we consider an alternative and arbitrary weighting scheme and report a moderate impact on both the confidence intervals and the relative importance of the SE and NSE components.

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 clean identification strategies based on popular techniques such as difference-in-differences, instrumental variables, and regression discontinuity design (Heath et al., 2023; Roberts and Whited, 2013). They have also accounted for multiple hypothesis testing and false discoveries to ensure that statistically significant results are not merely due to chance, thereby improving the reliability of inferences drawn from empirical analyses (Barras et al., 2010; Harvey et al., 2016; Harvey and Liu, 2020; Chordia et al., 2020; Chen, 2025).

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 (McLean and Pontiff, 2016; Harvey et al., 2016; Hou et al., 2020). In contrast, Jensen et al. (2023); Chen and Zimmermann (2022b) successfully replicated the findings of a large number of market anomalies papers, which aligns with recent evidence that replication rates in economics have improved (Brodeur et al., 2024).

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, data source, estimator, etc., multi-design studies aim to internalize methodological uncertainty by systematically spanning a range of protocol choices. We provide in the next section, an exhaustive survey of the multi-design literature in finance.

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 applicability of multi-design approaches throughout the discipline. The table provides the number of methodological decisions that are considered, as well as the total number of considered paths. We see that the listed studies employ various tools to summarize the large number of generated results, including plots of outcome distributions (e.g., box plots) and sensitivity analyses with respect to specific forks.

Moreover, since the pioneering work of Menkveld et al. (2024), it has become current practice to report so-called *nonstandard errors*. The latter aims 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). 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 if individual path outcomes differ from the median across all paths. With regard to the evaluation of SE, all contributions proceed as follows. They run each path, resulting in time-series of returns for each anomaly. Then, they employ bootstrapping techniques on these series to generate new average returns. Sampling uncertainty (i.e., the SE) then follows from the variations of these new returns across bootstrapped samples. In contrast, we propose in this paper to use resampling *before* running the paths.

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. (2025)	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 display 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 posterior 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.

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. Following Chen and Zimmermann (2022a), we propose a brief anatomy of choices made in the most influential papers in the field. We focus on US equities because they are the assets for which data and results are the most abundant. There are 331 anomalies in the latest version (as of January 2025) 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 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.
- **Holding period**: How long is the long-short portfolio 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).
- **Sample period**: The starting month for accounting data is most often taken to be June (190 instances) or December (54).

[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)).

- **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 0.30, 0.25, or 0.50 levels.
- **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.

This brief overview shows that the few commonly used options in the literature result in a wide array of possible choices. As an illustration, we see in Figure 1 that eight decisions lead to 576 possible paths. With the exception of two paths (the blue and orange ones), any other pair of paths will have one, some, or many steps in common.

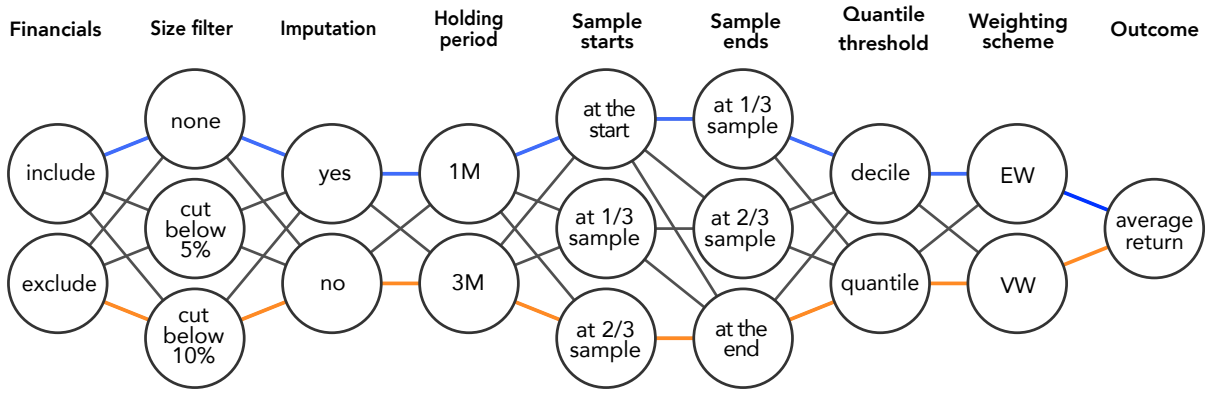

Figure 1: **Forking paths**. This figure displays the eight steps of the protocol along with the associated 576 paths. The ones in blue and orange follow entirely different steps.

2.3 Notations and definitions

We assume that the empirical part of any research process starts with a given dataset, which we call $\mathbb{D}$. Any 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 real number (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 $J = 8$ 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.

As any output is always associated to 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 $\hat{b}_q(\mathbb{D})$, allowing us to compute the sample correlation between the two series. Intuitively, $\rho_{p,q}$ measures how similar two paths are, with greater overlap leading to higher correlation and less overlap 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 protocols is to produce multiple estimates in order to gather evidence about an effect of interest. A natural approach is to summarize the resulting estimates using means, medians, box plots, etc. This method 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 the one derived from the sampled paths. Notably, Theorem 1 in Azriel and Schwartzman (2015) provides an upper bound for 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:

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

where $C > 0$ is some constant which can be taken to be equal to $C = 1/2$ for simplicity, and the norm is $\|\mathbf{\Sigma}_P\|_1 = P^{-2}\sum_{1\leqslant p,q\leqslant P}|\rho_{p,q}|$. The above result simply states that the error that the analyst makes when confusing the empirical and true cumulative distribution functions boils down to 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 this theoretical bound, we simulate $P = 3,000$ outcomes using a multivariate Gaussian distribution and use a Toeplitz matrix to generate 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. We consider two cases:

$$\gamma = 0.996 \text{ (low correlation), which corresponds to } \|\mathbf{\Sigma}_P\|_1 = 0.1525, \tag{4}$$

$$\gamma = 0.998 \text{ (high correlation), which corresponds to } \|\mathbf{\Sigma}_P\|_1 = 0.2776. \tag{5}$$

As shown in the left panel of Figure 2, we end up with two reasonable distributions displaying a large proportion of small correlations. In the high-correlation case, 64% of the

correlations are below 0.3 whereas in the low-correlation case, 81% of the correlations are below 0.3.

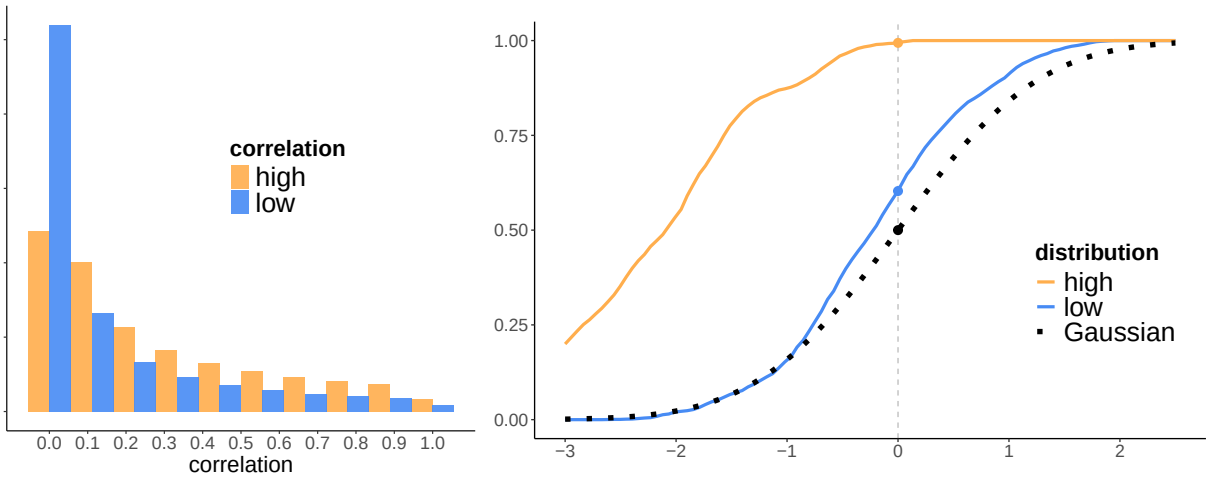

Figure 2: **Simulation exercise**. In the left panel, we plot the distributions of correlations in simulated samples in which the correlation level is set either high (orange) or low (blue). The simulation is based on Toeplitz matrices with ascending and descending diagonals equal to γ^n with $n = 0$ being the diagonal and $\gamma = 0.996$ (yielding $\|\mathbf{\Sigma}_P\|_1 = 0.1525$) and $\gamma = 0.998$ (corresponding to $\|\mathbf{\Sigma}_P\|_1 = 0.2776$), with the size of the matrix equal to $P = 3,000$. In the right panel, we plot the cumulative distribution function of a unique draw of the P variables. The black points mark the Gaussian density.

In the right panel of Figure 2, we plot the cumulative distribution function of a random draw of the corresponding multivariate Gaussian law in each sample. The proportion of observations smaller than zero is equal to 60% in the low-correlation case and 99% in the high-correlation case. This means that with low correlation, the error is equal to 10% and with high correlation to 49%. Squaring these errors yields respectively 1% and 24%, both of which fall below the bounds specified in Equation (3) and displayed in Equations (4) and (5).

This simple example shows that even with moderate correlations between outcomes, the distance separating the empirical distribution and the true one can be substantial. 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 asset pricing anomalies: the market capitalization (size factor), the book-to-market ratio (value factor), the past 12-to-1 month return (momentum factor), and the asset growth (investment factor). To conduct our tests, we extract data from Chen and Zimmermann (2022a)'s website over the 1951-2022 period.

For each sorting variable, 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 size, without replacement. 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 call 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). This highlights the potentially large errors that could occur when using the empirical distribution of effects (here, the average returns of sorted portfolios) as a proxy for the true distribution.

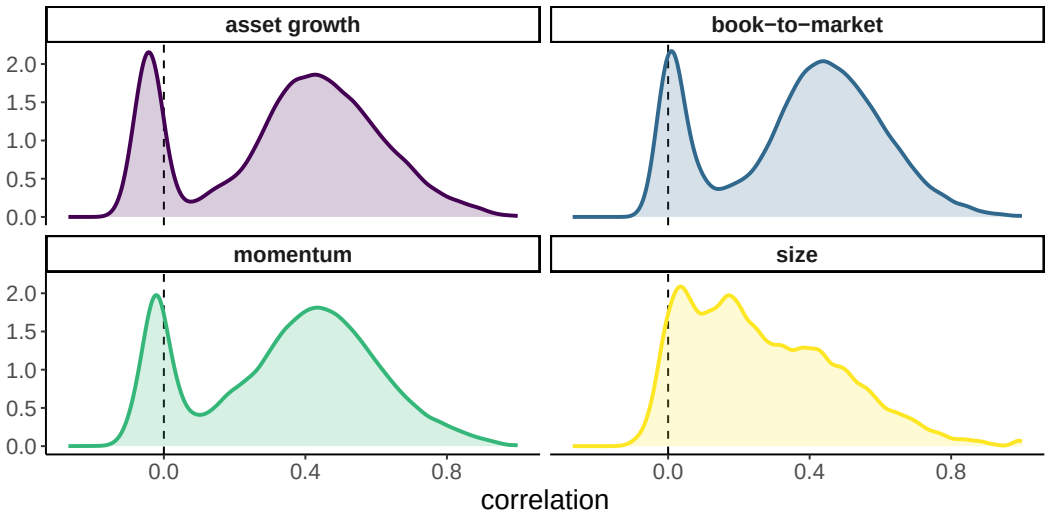

Figure 3: **Distribution of correlations**. We show the distribution of the correlations $\mathbb{C}or(\hat{b}_p, \hat{b}_q)$ for each of the four sorting variables, coded with colors. Correlations are computed on 500 random subsamples with a number of rows equal to 40% of that of the original dataset.

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 true mean μ_b, we must quantify the sample mean estimator:

$$\hat{\mu}_b = \frac{1}{P}\sum_{p=1}^{P} \hat{b}_p. \tag{6}$$

and its variance:

$$\sigma^2_{\hat{\mu}_b} = \mathbb{V}\left[\hat{\mu}_b\right] = \frac{1}{P^2}\sum_{1\leqslant p,q\leqslant P} \mathbb{E}\left[\left(\hat{b}_p - \mu_b\right)\left(\hat{b}_q - \mu_b\right)\right] = \frac{\sigma_b^2}{P^2}\sum_{p,q}\rho_{p,q}. \tag{7}$$

	Factors			
PANEL A: Common resampling	Asset growth	Book-to-market	Momentum	Size
$\|\hat{\mathbf{\Sigma}}_P\|_1 = P^{-2}\sum_{p,q}\|\hat{\rho}_{p,q}\|$	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				
$\|\hat{\mathbf{\Sigma}}_P\|_1 = P^{-2}\sum_{p,q}\|\hat{\rho}_{p,q}\|$	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 (7). 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 a specific dataset).

This 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 will be formally shown in the next section.

Second, the intervals will be tighter and more appealing if these correlations are small and/or symmetric around zero. In particular, it can be misleading to assume that the above variance is equal to the sample variance obtained from the paths, $\hat{\sigma}^2_b$. Indeed, this would only be true if the sum of non-diagonal correlations was null. However, this is unlikely to be true in practice, as 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 Equations (7), range between 0.26 and 0.37. These high levels imply large variances for $\hat{\mu}_b$.

We will show in Section 4 how 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. Reversely, if researchers follow paths that rarely or never overlap, then the correlation should be much lower.

As this is a testable assumption, we propose to examine it empirically. To do so, we define $d(p, q)$ as the number of choices that differ between path p and path q.[4] 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

[4]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 with path p passes for layer j.

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.

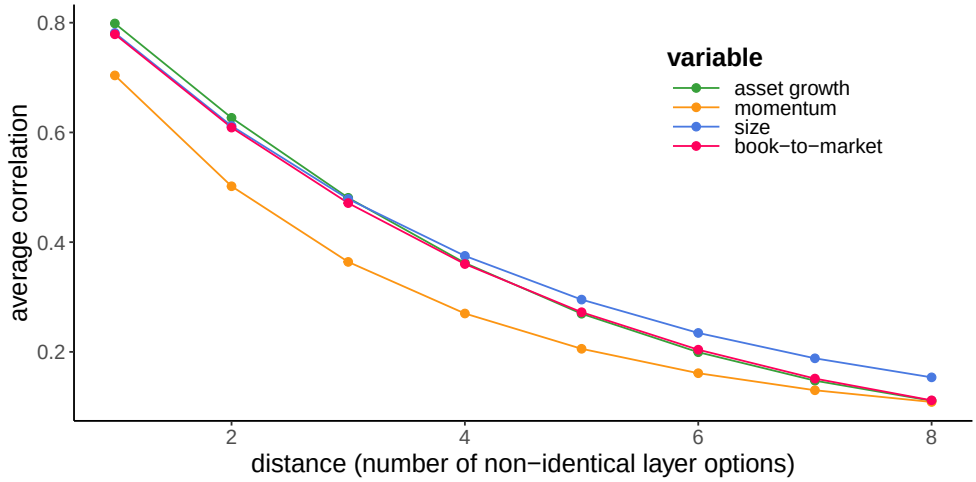

Figure 4: **Average correlation as a function of path distance**. We plot the mean correlation across all pairs of paths, as a function of the number of different choices between two paths (the distance between paths).

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 in 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 alter the distribution of correlations. Figure 5 displays the empirical density obtained by sub-sampling the data before each new iteration of the protocol. We clearly see that the 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. As shown in Panel B of Table 2, the norm of the correlation matrix in this case is close to 0.038 for all four factors, which is almost a tenfold reduction, compared to the case when we use the same dataset for all the paths (see Panel A).

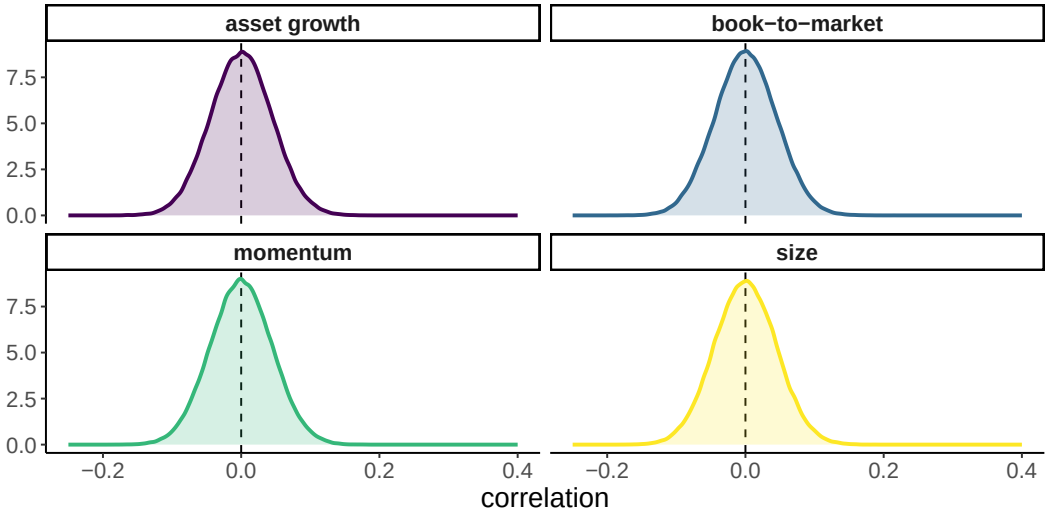

Figure 5: **Distribution of correlations after resampling**. We show the distribution of the correlations $\mathbb{C}or(\hat{b}_p, \hat{b}_q)$ for each of the four sorting variables, coded with colors. In this case, the data is subsampled before running each path and the number of rows corresponds to 40% of the size of the initial sample. Correlations are computed on 500 random subsamples.

4.2 Inference on the mean

In this section, we lay out ways to carry out inference on paths-generated outcomes. We are first interested in the empirical mean effect, $\hat{\mu}_b$ defined in Equation (6), which will naturally serve as proxy for the true mean. In order to proceed with inference, we must make some hypotheses on the underlying effect and how it is characterized by paths. We lay out a first assumption on which we will rely for the remainder of the paper.

Assumption 1. *The estimated effect of any path $\hat{b}_p$ follows the same law as the true effect, b. Moreover, this law is (i) unimodal, (ii) symmetric around its mode and mean μ_b, and (iii) has a finite variance, σ_b^2.*

This hypothesis is critical in the sense that if we do not rely on it, then inference based on path outcomes becomes almost impossible. Henceforth, we seek to build confidence intervals of the following form:

$$\text{IC}_\alpha = [\hat{\mu}_b - \Delta_\alpha, \hat{\mu}_b + \Delta_\alpha], \tag{8}$$

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

$$\mathbb{P}[|\mu_b - \hat{\mu}_b| \leqslant \Delta_\alpha] \geqslant 1 - \alpha. \tag{9}$$

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).[5] It states that, for a unimodal random variable Z symmetric around its mean and with variance σ_Z^2,

[5] An alternative approach would be to rely on the Central Limit Theorem (CLT). In most cases, variables

we have:

$$\mathbb{P}[|Z - \mathbb{E}[Z]| \leqslant v] \geqslant 1 - \left(\frac{2\sigma_Z}{3v}\right)^2 \tag{10}$$

for $v \geqslant 2\sigma_Z/\sqrt{3}$. Applied to the sample mean, this yields:

$$\mathbb{P}\left[|\hat{\mu}_b - \mu_b| \leqslant \frac{2\sigma_{\hat{\mu}_b}}{3\sqrt{\alpha}}\right] \geqslant 1 - \alpha, \tag{11}$$

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 (7). Intuitively, when $\sigma_{\hat{\mu}_b}$ decreases, the confidence interval shrinks, making it more informative. Conversely, as α shrinks 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 fact that we have to estimate the correlations, $\hat{\rho}_{p,q}$, based on N samples, which generates some additional uncertainty. We postpone the technical discussion of this particular point to Appendix A, and we simply formulate that, with at least probability $1 - \alpha$, the following upper bound holds for $\sigma^2_{\hat{\mu}_b}$:

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

where σ^2_* is an upper bound for σ^2_b (see Appendix B for a discusion on this bound). Plugging this into Equation (11), we get that the width of the confidence interval:

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

In Figure 6, we plot the confidence intervals for the four asset pricing anomalies. We consider three cases. First, we suppose that correlations are evaluated on paths run on identical samples (*common resampling*), which corresponds to the distributions shown in Figure 3. The second case pertains to when new samples are drawn prior to running a given path (*specific resampling*), hence the distributions of correlation 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) is equal to $1/P$. Finally, in the third case, we still implement specific resampling but we also assume ex-ante that the distribution is symmetric around zero (*specific resampling + symmetry*). This has two implications: first, the sum of correlations is equal to zero and second, there is no estimation error anymore (i.e.,

are assumed to be iid, but the CLT also holds under milder assumptions (see, e.g., White (2001)). The most general conditions under which the CLT holds are summarized in Jirak (2023). These conditions 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, these conditions do not apply and we cannot resort to the CLT.

$N = \infty$). Given Equation (12), the width of the interval reduces to

$$\Delta_\alpha = \frac{2\sigma_*}{3\sqrt{\alpha P}}. \tag{13}$$

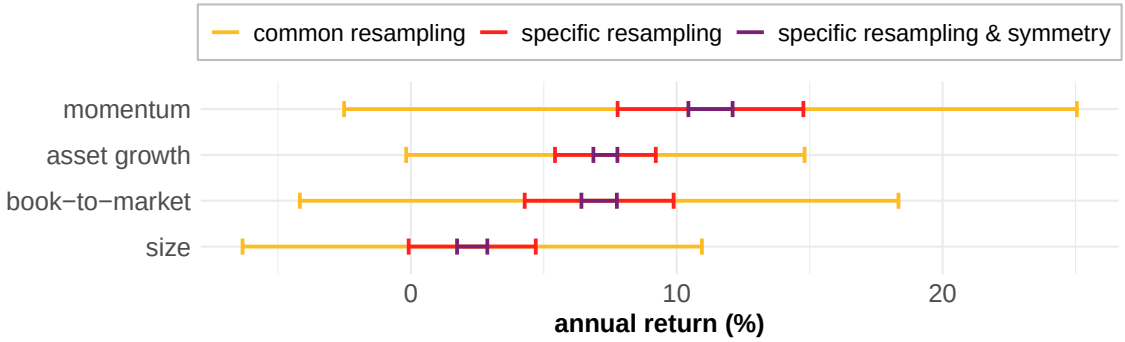

Figure 6: **Inference on the mean**. We plot, for $\alpha = 0.05$, the confidence intervals (8) of the mean of long-short returns in three situations: (i) common resampling corresponding to Figure 3 in **yellow**, (ii) path-specific resampling with estimation error in **red**) and (iii) path-specific resampling without estimation error ($N = \infty$), in **dark**. 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 derivations, σ_* is computed as follows. First, for each of the $N = 500$ samples, we compute the standard deviation of effects across paths, which gives 500 estimates for σ_b^2. Next, to be as conservative as possible, we take the maximum of these values. Finally, we take the upper bound σ_*^2 defined in Equation (29) in Appendix B.

Clearly, the ranges of the intervals in Figure 6 demonstrate the compelling 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 dark 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 that is at least tenfold 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:

$$\sigma^2_{\hat{\mu}_b} = \frac{\sigma_b^2}{P^2} \sum_{p,q} \rho_{p,q}. \tag{14}$$

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 posterior to spanning the paths.

Differently, 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 a specific one. 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 evaluate jointly both the standard and nonstandard errors. This common estimation framework leads to a single decomposition of the variance of the mean 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 recall below for two random variables X and Y with finite variance (Theorem 9.5.4 in Blitzstein and Hwang (2019)):

$$\mathbb{V}[Y] = \mathbb{V}[\mathbb{E}[Y|X]] + \mathbb{E}[\mathbb{V}[Y|X]]. \tag{15}$$

We seek to decompose the variance of the effect $\hat{b}_p(\mathbb{D})$, and, to this end, we can either condition according to samples, or to paths. For instance, conditioning with respect to samples leads to:

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

In the above expression, we notice that, in the second term, we recover the NSE that corresponds to the variance across paths, only, for one given dataset $\mathbb{D}$. What this version of the law of total variance shows is that this variance should be averaged across several $\mathbb{D}$. The first term, which corresponds to the variance of averages with respect to each $\mathbb{D}$ is, to the best of our knowledge, never 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:

$$\sigma_b^2 = \mathbb{V}\left[\hat{b}_p(\mathbb{D})\right] = \underbrace{\mathbb{V}\left[\underbrace{\mathbb{E}[\hat{b}_p(\mathbb{D})|p]}_{\text{avg effect across samples}}\right]}_{\text{variance of avg effect across paths}} + \underbrace{\mathbb{E}\left[\underbrace{\mathbb{V}[\hat{b}_p(\mathbb{D})|p]}_{\text{variance across samples}}\right]}_{\substack{\text{average variance across paths} \\ \text{SE reported in literature}}} \tag{17}$$

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 measured as the variance of the outcomes across paths. In Equation (17), 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 (17) and (16) (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:

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

$$\text{NSE} = \sqrt{\frac{\mathbb{V}[\mathbb{E}[\hat{b}|p]] + \mathbb{E}[\mathbb{V}[\hat{b}|\mathbb{D}]]}{2}}. \tag{19}$$

Crucially, in contrast with the conventions previously used in the literature, these identities verify:

$$\sigma_b^2 = \text{SE}^2 + \text{NSE}^2. \tag{20}$$

This decomposition expresses the total variance of the effects as the sum of two components: one arising from from sampling variability, and the other from the variability across paths. It is reasonable to question how the SE defined in Equation (18) 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 by 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 (18). We see that (1) the outcomes are quite consistent across anomalies and that (2) our approach produces SE that

are approximately half the size of those typically reported in the literature.

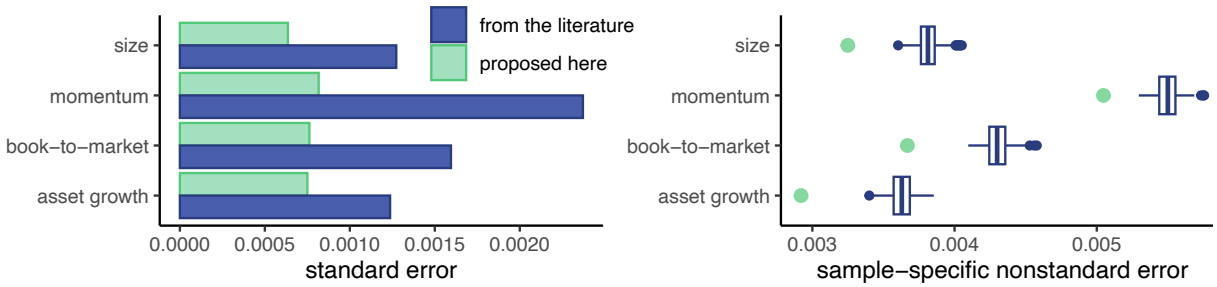

Figure 7: **Comparing SE and NSE methods**. In the left plot, we report the standard error for both our method (Equation (18)) and from the bootstrapping method suggested in the literature. In the right plot, we show the distribution of the NSE as it is computed in the literature (standard deviation 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 (19).

Finally, one can also wonder how the NSE defined in (19) 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 (17). In the right panel, we also feature the NSE as defined in Equation (19) 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 (19)) is larger than the path-focused component. Hence, upon averaging, the composite NSE is slightly lower, compared to the single-sample NSE. The relative importance of SE and NSE in Equation (20) will be further investigated in Section 5.2 below.

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 (20) into Equation (14), we can re-write the variance of the estimator of the mean:

$$\sigma^2_{\hat{\mu}_b} = \left(\text{SE}^2 + \text{NSE}^2\right) \sum_{p,q} \frac{\rho_{p,q}}{P^2}. \tag{21}$$

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 dominate the SE. Consequently, the NSE becomes the main determinant of the width of the confidence interval of 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 larger number of anomalies. We consider 33 factors out of the hundreds reported in the literature. We rely on the Open Source Asset Pricing dataset of Chen and Zimmermann (2022a), and we only keep the sorting variables that satisfy the three following 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 1951. This is because we cometimes 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 step (weighting scheme) 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 factor or anomaly reported by Chen and Zimmermann (2022a). In Figure 9, we report the confidence intervals defined by (12). We omit the intervals from common resampling because it 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 all of them do not overlap with zero. We also find some rare cases for which our results are more favorable: this occurs when the original returns are to the left of the red intervals (*High52*, *CBOperProf*, *OperProfRD*, *GP*). There are also many occurrences when original results are much to the right of our intervals (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 essentially from the cross-path dispersion of outcomes. Narrow intervals signal that variables can sustain a lot of methodological changes with limited change 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 found significant upon specific resampling, i.e., for which the red interval encompasses zero.

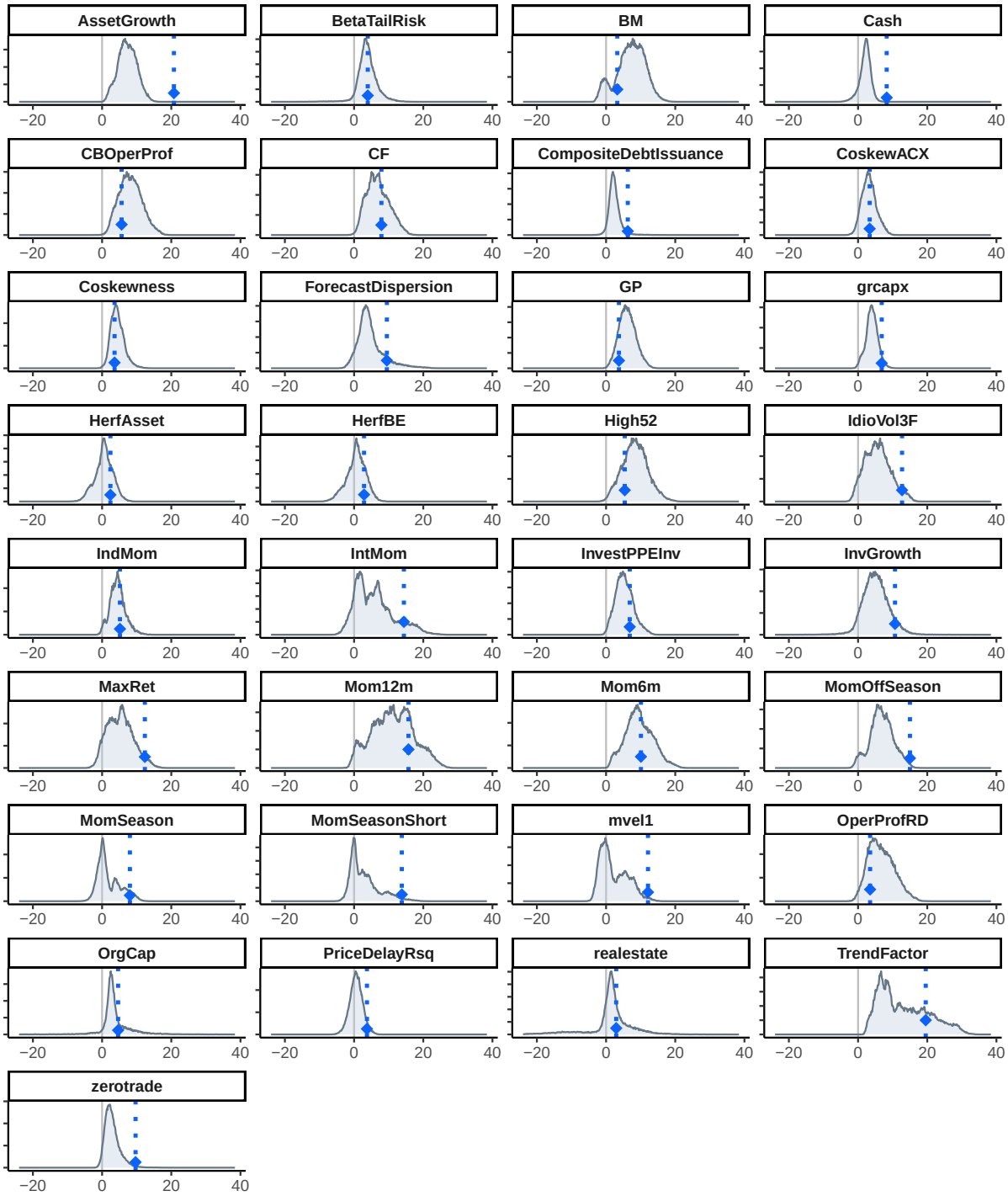

Figure 8: **Distribution of effects (average returns)**. We plot the distribution of average annual returns (in percents) across all 576 paths, bootstrapped samples, and sampling schemes. The names of characteristics are those of Chen and Zimmermann (2022a). The blue diamonds and the vertical points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

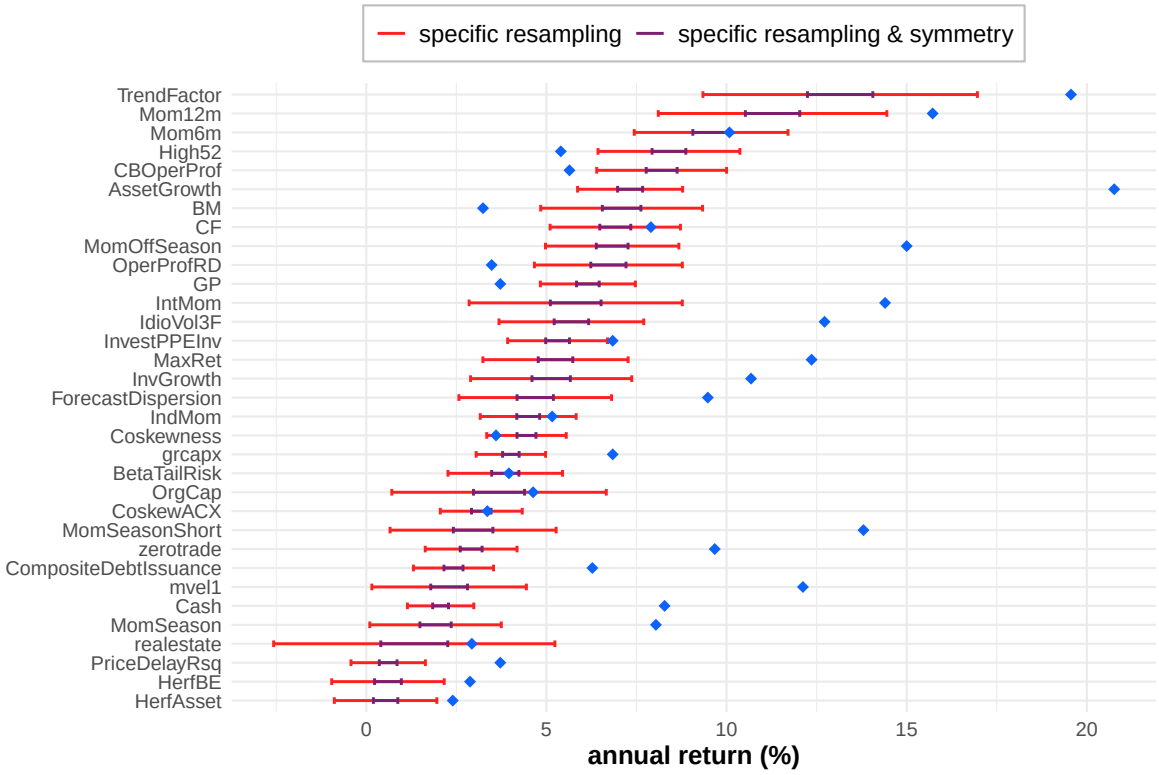

Figure 9: **Resilient anomalies**. We plot, for $\alpha = 0.05$, the confidence intervals (8) of the mean of long-short returns in two situations: (i) path-specific resampling with estimation error in **red**) and (ii) path-specific resampling without estimation error ($N = \infty$), in **dark**. By definition, the sample means lie in the middle of the intervals. The width of intervals is given by Δ_α in Equation (12). The blue diamonds locate the average return in the original studies.

5.2 Standard vs. nonstandard errors

We implement the variance decomposition displayed in Equations (18)-(20) for our 33 anomalies. 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.

Our empirical evidence is in line with the findings reported by Menkveld et al. (2024). We show in Figure 10 that variation in methodologies have a strong impact on the final

estimate, as shown by the large NSE reported in the figure. 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 (e.g. Soebhag et al. (2024)) and shows the importance of deriving both SE and NSE in a common framework.[6]

Our results about the magnitude and relative importance of NSE have important implication 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, for which we derived the following equation: $\sigma^2_{\hat{\mu}_b} = (\text{SE}^2 + \text{NSE}^2) \sum_{p,q} \rho_{p,q}/P^2$. 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 it is feasible, that is, not too costly. This is usually the case in empirical asset pricing.

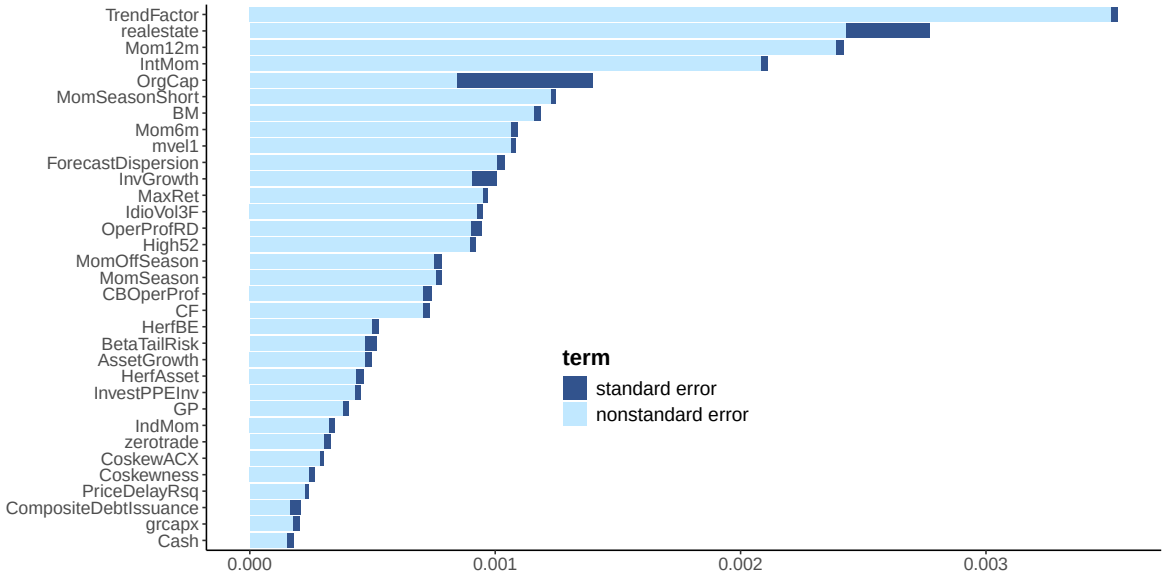

Figure 10: **Variance decomposition: standard versus nonstandard errors**. We display the decomposition of the variance in average returns proposed in Equations (18)-(20). The names of characteristics are those of Chen and Zimmermann (2022a).

6 Extension to non-uniform importance of paths

Thus far, we have described an agnostic approach that treats all outcomes as equally important. However, alternative weighting schemes. 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.

[6] As shown in Figure 7, when implementing the methodology used in the literature, we find a NSE-to-SE ratio between 2 and 3, which is quite in line with the results in Soebhag et al. (2024)).

In this case, the sample mean estimator is written:

$$\hat{\mu}_b = \sum_{p=1}^{P} \omega_p \hat{b}_p, \quad \text{with} \quad \sum_{p=1}^{P} \omega_p = 1. \tag{22}$$

and its variance:

$$\sigma^2_{\hat{\mu}_b} = \mathbb{V}\left[\hat{\mu}_b\right] = \sigma^2_b \sum_{p,q} \omega_p \omega_q \rho_{p,q}. \tag{23}$$

where we recover (6) and (7) by taking $\omega_p = 1/N$. Moreover, the sample standard deviation of effects is:

$$\hat{\sigma}^2_b = \frac{1}{P-1} \sum_{p=1}^{P} \omega_p (\hat{b}_p - \hat{\mu}_b)^2. \tag{24}$$

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 which 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:

$$\omega_p = \frac{s_p}{\sum_{p=1}^{P} s_p}. \tag{25}$$

In Figure 11 below, we reproduce the analysis from Figure 9 but with the weighted averages and variances defined in Equations (22)-(24) with weights equal to (25). We are only interested in the situations with path-specific resampling. In this case, because of the symmetry in correlations, the term in (23) will not move much compared to the baseline situation, and it is (24) that will drive the changes in variability.

From afar, the confidence intervals are similar to those of Figure 9, with minor differences. The range of the intervals are roughly unchanged, hence it is mostly on the mean that the weighting has the most influence. The most important shift is perhaps that of *IndMom*, with an interval that is now close to crossing the origin. Nevertheless, the same four factors are found to be sensitive to methodological changes. This indicates that 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 mostly one notable difference, compared to uniform weights (Figure 10): the ordering is not exactly the same. For instance the *realestate* variable ranks fourth in the new plot, whereas it was second in the original one. 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 standard errors versus nonstandard ones.

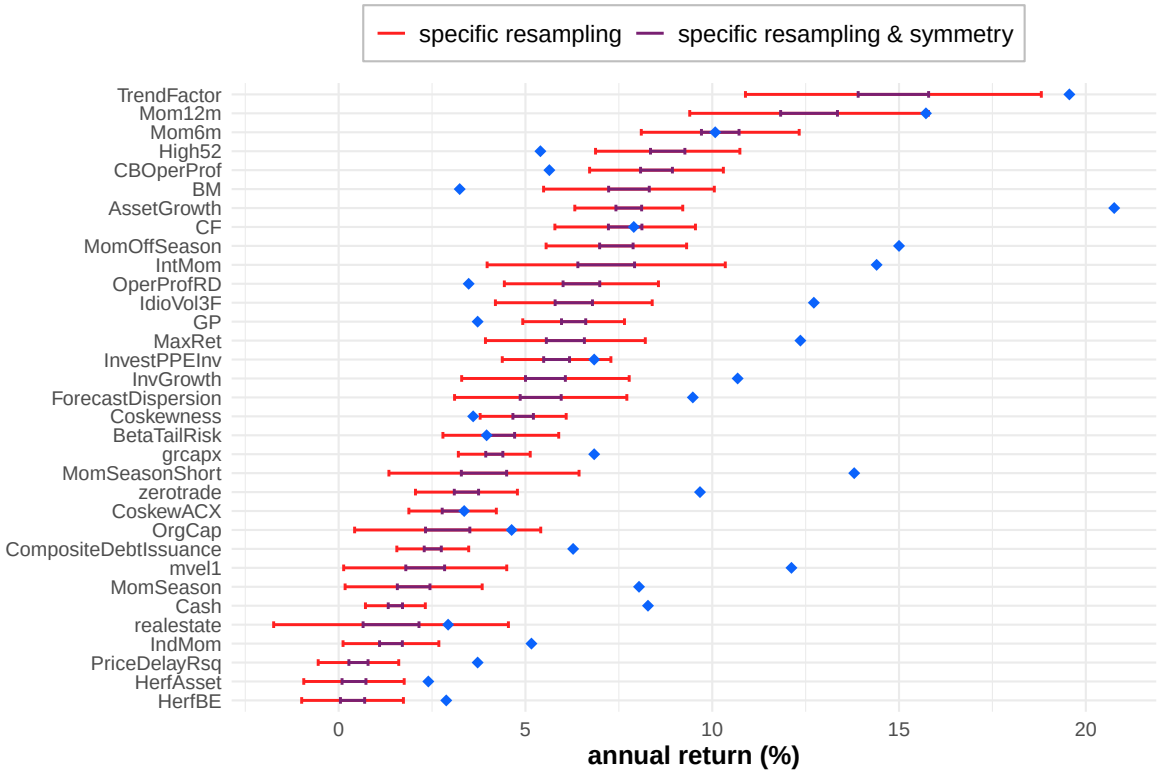

Figure 11: **Resilient anomalies with non-uniform weights**. We replicate the analysis in Figure 9 but with non-uniform weights. The baseline path is the blue one in Figure 1 and each path has a weight proportional to 0.75^d, where d is the distance to the baseline path (see Equation (25)). Averages and variances of outcomes are computed according to Equations (22) and (23).

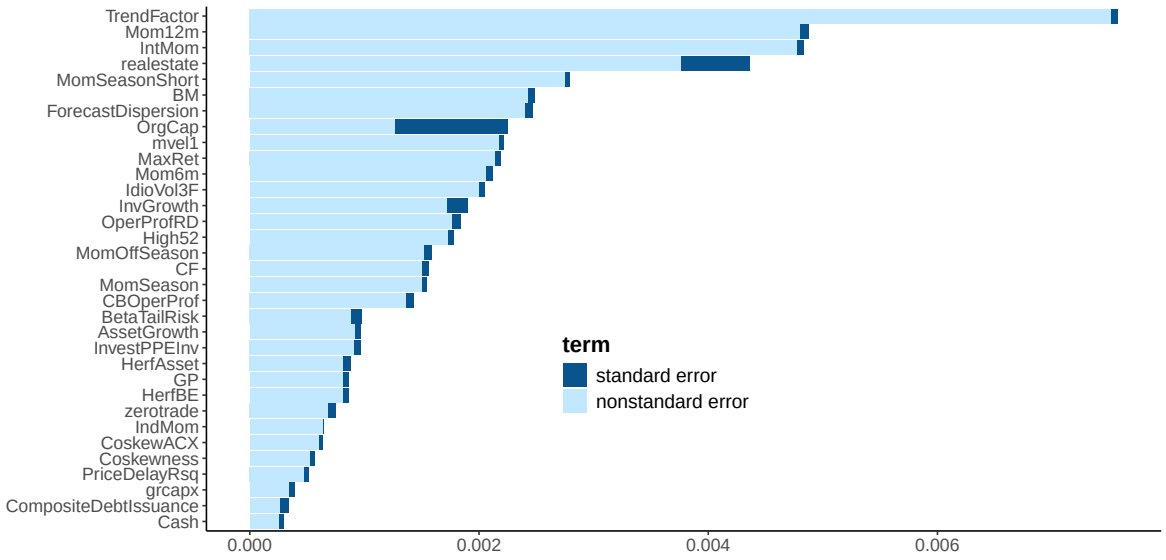

Figure 12: **Variance decomposition under non-uniform weights**. We display the decomposition of the variance in average returns proposed in Equation (20) but with weights given in Equation (25). The names of characteristics are those of Chen and Zimmermann (2022a).

7 Conclusion

Menkveld et al. (2024) is a truly *path-breaking* paper in finance, 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 effect. Specifically, we derive a canonical decomposition of the variance of the mean effect. 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 all correlations 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 variations of the effect that are due to sampling noise. Finally, the third component is the nonstandard error (NSE) that results from the uncertainty generated from 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 to shrink the width of confidence intervals by a factor three. Moreover, assuming a symmetric distribution of correlation 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

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

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

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

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

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

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

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

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

Brodeur, A., D. Mikola, and N. Cook (2024). Mass reproducibility and replicability: A new hope.

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

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

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

Chen, A. Y. (2025). Do t-statistic hurdles need to be raised? *Management Science 0*(0), null.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Jensen, T. I., B. T. Kelly, and L. H. Pedersen (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.

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

Menkveld, A., A. Dreber, F. Holzmeister, M. Johannesson, J. Huber, M. Kirchler, S. Neususs, M. Razen, U. Weitzel, 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)*.

Olkin, I. and J. W. Pratt (1958). Unbiased estimation of certain correlation coefficients. *Annals of Mathematical Statistics*, 201–211.

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

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

Silberzahn, R., E. L. Uhlmann, D. P. Martin, P. Anselmi, F. Aust, E. Awtrey, Š. Bahník, F. Bai, C. Bannard, E. Bonnier, 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., B. van Vliet, and P. Verwijmeren (2024). Non-standard errors in asset pricing: Mind your sorts. *Journal of Empirical Finance 78*, 101517.

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

White, H. (2001). *Asymptotic theory for econometricians*. Academic Press.

Zhang, M., T. Lu, and C. Shi (2025). *Navigating the Factor Zoo: The Science of Quantitative Investing*. Taylor & Francis.

A Inference on the mean

Let us recall Equation (11):

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

which relies on $\sigma_{\hat{\mu}_b}$, and this value depends mostly on the information on the correlations $\rho_{p,q}$. There are two natural estimators for the correlations, $\hat{\rho}_{p,q}$: the Olkin and Pratt (1958) estimator[7] and the sample correlation, for which there are many estimators of the variance. Having tested both for $N = 100$ and $N = 500$, we obtain almost indistinguishable results.

Gnambs (2023) suggests that a reasonable choice for the variance of the estimator 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

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

We will henceforth assume that this inequality holds: for $N \geqslant 100$, simulation studies (Gnambs (2023)) suggest that the error on the standard error is marginal.

The Bienaymé-Chevbyschev inequality, applied to estimated correlation, then implies,

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

where we have simply used Equation (26) and assumed that the estimator $\hat{\rho}_{p,q}$ is unbiased. Hence, for N large enough, the estimations will be accurate enough. Then, with probability $1 - \alpha$ at least,

$$|\rho_{p,q} - \hat{\rho}_{p,q}| \leqslant \frac{2}{3\sqrt{\alpha N}}, \tag{27}$$

and

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

where we recall that, by Assumption 1, σ_*^2 is a known upper bound for σ_b^2. Plugging this in Equation (11), we finally have

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

which is the sought interval.

[7] In this case, $\hat{\rho} = \hat{r}\left(1 - \frac{1-\hat{r}^2}{N-3}\right)$, where $\hat{r}$ is the sample correlation.

B Discussion on σ_*^2

The true variance of b, σ_b^2 is unknown and estimated with $\hat{\sigma}_b^2$. 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 additional assumptions that allow to obtain a bound on σ_b^2 with a confidence level of $1-\alpha$.

Suppose we have an unbiased estimation σ_b^2 from $P(P-1)/2$ 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 below, we thus have

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

The bound in the bracket depends on σ_b, which is unknown, but again considering the two cases $\sigma_b^2 < \hat{\sigma}_b^2$ and $\sigma_b^2 > \hat{\sigma}_b^2$, we are able to obtain

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

and hence we can set

$$\sigma_*^2 = \hat{\sigma}_b^2\left(1 + \frac{1}{3\sqrt{\alpha(P-1)}/(2\hat{\sigma}_b^2) - 1}\right). \tag{29}$$

Assumption 2. *It holds that:*

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

The last two points seem reasonable when data is reshuffled prior to each analysis. For example, we provide in Figure 13 the distribution of the correlation between squared effects. Interestingly, this distribution can also be fitted with a centered beta law with large $\mathfrak{a}$. We now proceed with a result on the variance of the sample variance.

Lemma 1. *Under Assumption 2, the variance of the sample variance is* $\mathbb{V}[\hat{\sigma}_b^2] = \frac{2\sigma_b^4}{P-1}$.

Proof. First, as a side note, let us note that

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

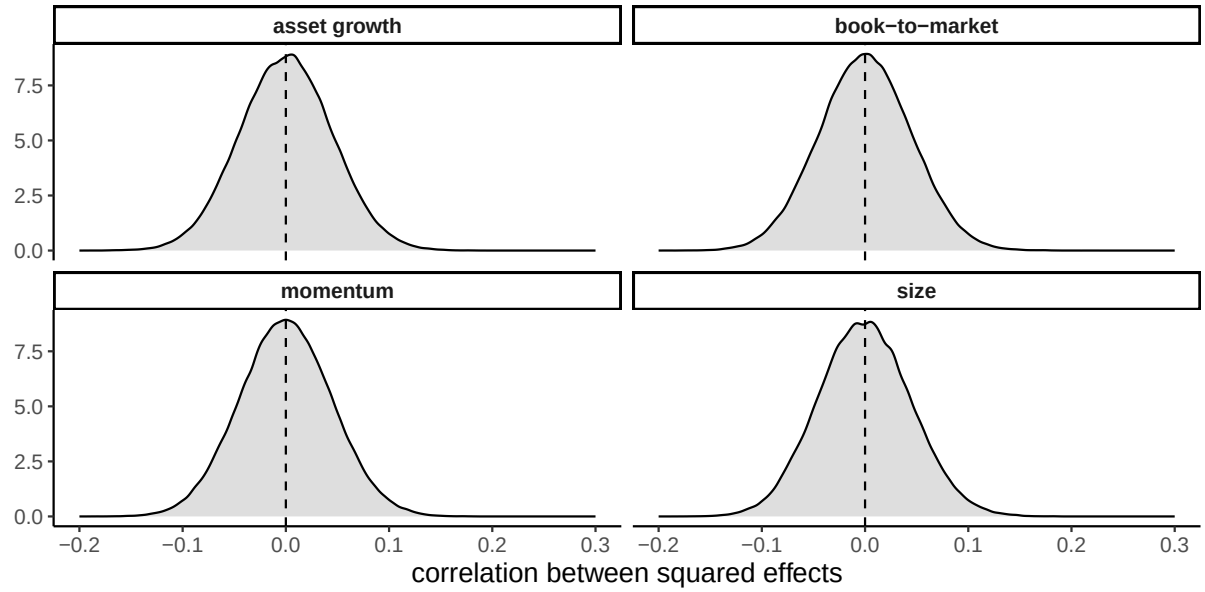

Figure 13: **Distribution of correlations between squared effects**. We show the distribution of the correlations $\rho = \mathbb{C}or(\hat{b}_p^2, \hat{b}_q^2)$ for each of the four sorting variables. Correlation are computed on $N = 500$ samples and the samples are generated separately for each path, following our path-specific approach.

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., $\mathbb{E}[\hat{\sigma}_b^2] = \sigma_b^2$.[8]

Moreover, by Assumption 2,

$$\mathbb{V}\left[\sum_{p=1}^{P}\hat{b}_p^2\right] = \sum_{p=1}^{P}\mathbb{V}\left[\hat{b}_p^2\right] + \underbrace{\sum_{p\neq r}\text{Cov}\left(\hat{b}_p^2, \hat{b}_r^2\right)}_{=0} = \sum_{p=1}^{P}\mathbb{E}\left[\hat{b}_p^4\right] - \mathbb{E}\left[\hat{b}_p^2\right]^2 \tag{30}$$

$$= P(3\sigma_b^4 + 6\sigma_b^2\mu_b^2 + \mu_b^4 - (\sigma_b^2 + \mu_b^2)^2) = P(2\sigma_b^4 + 4\sigma_b^2\mu_b^2) \tag{31}$$

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

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.[9] Finally, we will need the following expression

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

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

Now, aggregating everything in a bigger picture and aggregating the pieces:

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

□

[9] If the effects have zero mean, then Isserlis' theorem for $\mathbb{E}[\mu_p\mu_q\mu_r\mu_s]$ and Asssumption 2 imply:

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