Correlated False Discoveries and Asset Pricing Anomalies

Daniele Bianchi
Queen Mary, University of London
d.bianchi@qmul.ac.uk

Junpei Komiyama
NYU Stern School of Business
junpei.komiyama@nyu.edu

Kenichiro McAlinn
Temple University, Fox School of Business
kenichiro.mcalinn@temple.edu

Leyong Yang
NYU Stern School of Business
ly2143@stern.nyu.edu

Abstract

This paper introduces a novel dueling double-bootstrap approach for assessing the replicability of asset pricing anomalies and factors while controlling false discovery rate (FDR) under general correlation structures. We establish theoretical guarantees for FDR control and provide simulation evidence that our method outperforms widely used multiple testing approaches. Empirically, we demonstrate that accounting for unknown hypotheses correlation often increases replication rates across popular equity and bond portfolios, with variations by weighting schemes, risk adjustments, return calculations, and testing samples. Our findings underscore the importance of robust multiple testing frameworks in empirical asset pricing for identifying genuine portfolio anomalies and risk factors.

Keywords: Asset Pricing, Portfolio Anomalies, Risk Factors, Multiple Hypothesis Testing, False Discovery Rate, Bootstrap Methods.

JEL codes: G12, G14, C12, C15, C58

1 Introduction

In their influential work, Harvey et al. (2016) highlighted the difficulty researchers and practitioners face in identifying which of the many proposed portfolio "anomalies" and risk factors provide genuine insights into the cross-sectional variation of asset returns. The lack of rigorous research practices – such as appropriate corrections for multiple testing – has exacerbated issues like data snooping and p-hacking, leading to often spurious statistically significant results. Consequently, many reported discoveries fail to endure rigorous scrutiny, undermining both their academic evidence and practical relevance.

A central challenge emerges from the correlated nature of portfolio returns, stemming from overlapping economic theories and shared data inputs (e.g., Jensen et al., 2023). A simple correlation analysis in Section 4 highlights this issue, revealing a heterogeneous dependence structure among both equity and bond portfolio returns – both raw and market-risk-adjusted – regardless of the portfolio construction and return calculation approach. These correlations violate the independence assumption of most testing procedures, substantially complicating the evaluation of anomaly and risk factor replicability (e.g., Schwartzman and Lin, 2011).

Traditional multiple-hypothesis testing (MHT) approaches often rely on test independence – a condition rarely satisfied in asset pricing. When returns are correlated, these methods can produce misleading results, as shared characteristics rather than unique risk premium components may drive statistical significance.

To address these challenges, we develop a novel MHT framework that controls the false discovery rate (FDR) while remaining agnostic to the correlation structure of the underlying hypotheses. Our approach implements a dueling double-bootstrap (`DDB`) procedure that first estimates a lower bound on true null hypotheses before computing optimal p-value threshold corrections. We establish theoretical guarantees that our method controls the FDR with probability approaching one as the number of bootstrap samples increases, even when hypotheses exhibit arbitrary correlations.

Through an extensive simulation study, we demonstrate that our proposed method consistently outperforms widely used approaches for controlling FDR (e.g., Benjamini and Hochberg, 1995; Yekutieli and Benjamini, 1999; Benjamini and Yekutieli, 2001; Storey, 2002; Storey et al., 2004; Benjamini et al., 2006). Our `DDB` approach maintains robust FDR control while achieving higher rejection rates across various scenarios. The method's advantages are particularly evident under high correlation, where traditional methods such as Benjamini and Hochberg (1995) (`BH`) and Benjamini and Yekutieli (2001) (`BY`) often fail to adequately control FDR or exhibit low power. Compared to Storey-based approaches, `DDB` delivers superior FDR control while achieving comparable rejection rates.

We apply the dueling double-bootstrap (`DDB`) to examine the extent to which equity portfolio anomalies and bond risk factors proposed in the asset pricing literature can be replicated based on the last four decades of returns. Following Hou et al. (2020), we define a portfolio strategy as successfully replicated if the average raw or risk-adjusted return is statistically significant under FDR control.

Drawing from a comprehensive dataset of 153 portfolio returns by Jensen et al. (2023), we first assess the significance of equity portfolio anomalies across equal-weight, value-weight, and capped value-weight schemes, with the latter winsorizing market capitalizations at the 80th NYSE percentile. Our initial results reveal that replication success in equity anomalies depends crucially on portfolio construction methods. Market risk-adjusted returns (i.e., CAPM alphas) on equal-weight portfolios show markedly higher replication rates under `DDB` (87%) compared to value-weight portfolios (68%). This disparity reflects the tendency of value-weight portfolios to concentrate on large-cap stocks, potentially diluting underlying signals. The improved performance of capped value-weight portfolios (75% replication rate) supports this interpretation.

The results further demonstrate that the choice between raw and risk-adjusted returns also significantly impacts replication outcomes. Raw returns yield lower replication rates compared to CAPM alphas for equal-weight (75% vs 87%), value-weight (29% vs 68%), and capped value-

weight (48% vs 75%) portfolios, This is consistent with the argument in Jensen et al. (2023) that risk adjustment helps isolate anomalous components from broader market movements.

The importance of portfolio constructions is confirmed when using an alternative set of 176 characteristic-managed portfolios from Chen and Zimmermann (2021). The results confirm that replication rates are lower for raw returns compared to CAPM alphas, as well as lower for more restrictive MHT methods that fail to account for correlated hypotheses. Consistent with the main empirical analysis, the Storey method achieves the highest replication rates due to its extreme p-value thresholds; however, this advantage comes at the cost of compromised FDR control.

Expanding our analysis beyond equities, we examine the replication rates for 15 tradable bond portfolios from Dickerson et al. (2023a). The results depict a less optimistic view compared to equity portfolio anomalies. When considering duration-adjusted returns, the replication rate drop to 23% for raw returns and 13% for market risk-adjusted returns (CAPM alphas). In the latter case, only two factors show significant excess returns adjusted for market risk (CAPM alphas): a bond momentum factor constructed from equity momentum and a bond portfolio based on earnings announcement drift. This finding echoes Dickerson et al. (2023a), which show that earnings announcements represent a potentially strong source of priced risk in the cross-section of stocks and bonds.

When comparing our `DDB` approach to existing methods, notable differences emerge. Traditional MHT methods, such as `BH` and `BY`, yield overly conservative results by ignoring portfolio correlations. In contrast, Storey-based methods generate optimistic findings with high replication rates, but at the cost of compromised FDR control in dependent settings. By explicitly accounting for correlation structures, our dueling double-bootstrap approach achieves a more balanced assessment of anomaly replicability, avoiding both the overly restrictive and overly lenient extremes of existing methods.

Interestingly, contrary to equity portfolios, adjusting bond risk factors for their market exposure substantially decreases replicability. The replication rate of bond portfolios is reduced

by half when moving from raw returns to CAPM alphas. Nevertheless, we confirm that more restrictive multiple hypothesis testing (MHT) methods, which fail to account for correlated hypotheses, consistently produce lower replication rates.

The main empirical results are confirmed across different validation samples. Following Harvey and Liu (2021), we examine variations in sample length and cross-sectional coverage. Shorter samples naturally yield lower replication rates – due to reduced statistical power and diminished precision of parameter estimates (e.g., mean returns, standard deviations) – the relative rates between equal-weight, value-weight, and capped value-weight portfolios is consistent with the main empirical analysis.

Numerous studies have examined the robustness of empirical asset pricing findings to data mining (Barras et al., 2010; Green et al., 2013; Hou et al., 2015; Harvey et al., 2016; Yan and Zheng, 2017; Chordia et al., 2020; Harvey and Liu, 2020; Hou et al., 2020; Harvey and Liu, 2021; Chen and Zimmermann, 2022; Chen and Dim, 2023; Jensen et al., 2023). Hou et al. (2020) investigated anomaly robustness using alternative portfolio construction methods, presenting a pessimistic outlook for anomaly replicability without incorporating state-of-the-art multiple testing frameworks. This pessimistic view was later challenged by Jensen et al. (2023), who demonstrated higher replication rates using a parametric Bayesian MHT approach. However, most existing MHT approaches remain limited by their restrictive correlation assumptions, significantly constraining their practical applicability in realistic asset pricing settings.

We contribute to this literature by developing a dueling double-bootstrap approach that balances statistical power and FDR control under arbitrary correlation structures. Our approach bridges the gap between the overly conservative method of Benjamini and Hochberg (1995) and the overly lenient approach of Storey et al. (2004), providing a more robust framework for controlling FDR when assessing the significance of portfolio anomalies and risk factors. Most importantly, our findings confirm substantially higher replication rates than previously documented – at least for equity factors – while effectively addressing the critical issue of hypothesis dependencies.

2 Multiple hypothesis testing

Consider the $i \in [N] = \{1, 2, \ldots, N\}$ portfolio anomalies over $t \in [T]$ time steps. The returns $y_{i,t}$ are often modeled as:

$$y_{i,t} = a_i + b_i^\top \mathbf{x}_t + \varepsilon_{i,t}, \quad i = 1, \ldots, N, \tag{1}$$

where a_i represents the risk-adjusted return, $\mathbf{x}_t \in \mathbb{R}^p$ is a column vector of p risk factors, and $\varepsilon_{i,t}$ are residual errors satisfying $\mathbb{E}[\mathbf{x}_t \varepsilon_{i,t}] = \mathbf{0}$ $\forall i \in [N],\ t \in [T]$ and $\text{Cov}(\varepsilon_{i,t}, \varepsilon_{j,s}) = 0,\ \forall i, j \in [N],\ t \neq s$. The studentized t-statistics for the estimated $\widehat{a}_i$ in Eq.(1) take the form:

$$\widehat{\alpha}_i = \frac{\widehat{a}_i}{\widehat{\sigma}_i / \sqrt{T}}, \tag{2}$$

where $\widehat{\sigma}_i^2$ is an estimate of the sample error variance. The t-statistics in Eq.(2) follow a Student-t distribution with $T-p-1$ degrees of freedom, i.e., $\widehat{\alpha}_i \sim \mathcal{T}_{T-p-1}$ (e.g., Casella and Berger, 2024, Section 5). The p-values, denoted as $\mathbf{p} = (p_1, \ldots, p_N)$, quantify the probability of observing test statistics as extreme as or more extreme than $\widehat{\boldsymbol{\alpha}}$ under the null.[1]

Let $\mathcal{H} = [N]$ denote the set of all hypotheses, with $\mathcal{H}_0 = \{i \in [N] : \alpha_i = 0\}$ and $\mathcal{H}_1 = \mathcal{H} \backslash \mathcal{H}_0$ the null and alternative hypotheses, respectively. Define $N_0 = |\mathcal{H}_0|$ and $N_1 = |\mathcal{H}_1|$ as the number of null and alternative hypotheses, with proportions $\pi_0 = N_0/N$ and $\pi_1 = N_1/N$. A multiple hypothesis testing procedure is a function $R(\widehat{\boldsymbol{\alpha}}) : \mathbb{R}^N \to 2^{[N]}$ that maps the observed statistics $\widehat{\boldsymbol{\alpha}}$ to a subset of hypotheses that are likely to belong to $\mathcal{H}_1$. The i-th hypothesis is

[1] For a single hypothesis i, the p-value is defined using the tail area $p_i(\widehat{\alpha}_i) = 2\int_{|\widehat{\alpha}_i|}^{\infty} f(t)\,dt$ for a two-sided test and $p_i(\widehat{\alpha}_i) = \int_{\widehat{\alpha}_i}^{\infty} f(t)\,dt$ for a one-sided test, where $f(t)$ is the probability density function of the Student-t distribution.

rejected if $i \in R(\widehat{\boldsymbol{\alpha}})$, and the overall testing results can be classified as follows:

	$i \notin R(\widehat{\boldsymbol{\alpha}})$	$i \in R(\widehat{\boldsymbol{\alpha}})$	**Total**
$\mathcal{H}_0$	$N_{0\|0}$	$N_{1\|0}$	N_0
$\mathcal{H}_1$	$N_{0\|1}$	$N_{1\|1}$	N_1
Total	$N - R$	R	N

where $N_{1|0}$ are false positives (nulls incorrectly rejected, also known as type I errors), $N_{1|1}$ are true positives (alternatives correctly rejected), $N_{0|1}$ are false negatives (alternatives incorrectly retained, also known as type II errors), and $N_{0|0}$ are true negatives (nulls correctly retained).

The false discovery rate (FDR) is the expected proportion of false positives among all rejected hypotheses:

$$\text{FDR}(\boldsymbol{\alpha}) = \mathbb{E}_{\mathbf{u}\sim\mathcal{U}}\left[\frac{N_{1|0}}{\max(R,1)}\right], \tag{3}$$

where $R = |R(\widehat{\boldsymbol{\alpha}})| = N_{1|0} + N_{1|1}$, and $\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha} = \mathbf{u} \sim \mathcal{U}$ is the alpha estimation error.

The objective of this paper is to propose a method that maximizes statistical power (correctly identifying true positives) under an arbitrarily correlated set of hypotheses $\mathcal{U}$ while ensuring $\text{FDR}(\boldsymbol{\alpha}) \leq q$, for $q \in (0,1)$ and any $\boldsymbol{\alpha}$. In the following sections, we provide an overview of popular multiple hypothesis testing (MHT) methods for controlling the FDR. These methods serve as the foundation for developing and comparing our proposed approach.

2.1 Step-up methods

One of the most widely accepted methods for controlling the FDR was proposed by (Benjamini and Hochberg, 1995) (`BH`), which adopts a step-up procedure. The `BH` method sorts the p-values in ascending order, $p_1 \leq p_2 \leq \cdots \leq p_N$, and rejects hypotheses sequentially up to the largest index i satisfying,

$$p = \max_i \left\{ p_i : \frac{q \cdot i}{N} \geq p_i \right\}. \tag{4}$$

This method controls the FDR at level $\pi_0 q \leq q$ under the assumption of independent hypotheses.[2] For correlated hypotheses, a logarithmic correction is required. Specifically, replacing q with $q/(\log N + 0.5)$ in Eq. (4) guarantees FDR control (see Benjamini and Yekutieli, 2001). However, in our simulation study, we show that such correction is overly conservative, leading to lower statistical power compared to our dueling double-bootstrap approach.

For independent hypotheses, an optimal oracle method was proposed by Storey (2002). This method relies on the ratio of true discoveries, π_1, which is unknown *a priori*. To approximate π_1, a plug-in estimator is used:

$$\widehat{\pi}_1 = \max\left(0, \frac{\widehat{F}_{\widehat{\boldsymbol{\alpha}}}(p_0) - p_0}{1 - p_0}\right), \tag{5}$$

where $\widehat{F}_{\widehat{\boldsymbol{\alpha}}}(p_0) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[p_i \leq p_0]$ is the empirical cumulative distribution function (CDF) of the p-values, and $p_0 \in (0, 1)$ is a reference point. Benjamini et al. (2006) recommend setting $p_0 = 0.5$. Using $\widehat{\pi}_1$, the threshold for rejecting hypotheses is defined as:

$$p = \sup\left\{p_i : \frac{q \cdot i}{N\,(1 - \widehat{\pi}_1)} \geq p_i\right\} \tag{6}$$

While Eq. (6) controls the FDR under independent hypotheses, it fails to maintain FDR control in correlated settings (Romano et al., 2008).

Storey et al. (2004) extend this framework by selecting p_0 adaptively to balance the bias and variance of $\widehat{\pi}_1$ through bootstrapping p-values. This adaptive procedure asymptotically controls the FDR under weak dependence conditions, which include independence and mixing-type dependence. However, it does not account for arbitrary dependence structures, such as constant correlation across all p-values.[3]

[2] It also controls the FDR under a specific type of correlation called "positive regression dependence on a subset" (PRDS) condition. This condition ensures that increasing one test statistic does not decrease the likelihood of observing larger values of other test statistics within a specific subset. Formally, for a family of test statistics $(t_1, t_2, \ldots, t_m)$, PRDS requires that for any subset $S \subset \{1, 2, \ldots, m\}$, the conditional probability $P(t_i \geq \tau \mid t_S \geq \tau_S)$ is non-decreasing in τ for $i \notin S$, where t_S represents the test statistics in subset S.

[3] The method excludes cases where all p-values exhibit constant correlation, a scenario which we consider in the simulation study.

While Romano et al. (2008) established a framework for controlling asymptotic FDR under general correlations, a critical gap remains in their approach as they provide no explicit method for estimating $\widehat{\pi}_1$. Our simulation results reveal that standard estimators of $\widehat{\pi}_1$, including Eq. (5), break down under correlated hypotheses, typically yielding inflated estimates. Furthermore, Romano et al. (2008) leaves the practical implementation of their bootstrapping methodology unspecified. In Section 2.2, we demonstrate that conventional cross-sectional bootstrap (Fama and French, 2010) offers a robust solution for accurately estimating correlation structures.

2.2 Dueling double-bootstrap (DDB)

Since $\widehat{F}_{\widehat{\boldsymbol{\alpha}}}(p) = i/N$ if threshold is $p = p_i$, a step-up method can be expressed more generally as:[4]

$$p = \sup \left\{ p_i : \widehat{F}_{\widehat{\boldsymbol{\alpha}}}(p_i) \geq \frac{p_i}{c_q} \right\}, \tag{7}$$

where c_q is a scaling factor that depends on the FDR method. For example, $c_q = q$ corresponds to the method of Benjamini and Hochberg (1995), and $c_q = q/(1 - \widehat{\pi}_1)$ corresponds to the adaptive procedure proposed by Storey et al. (2004).

Under the assumption that all null hypotheses are independently distributed and the complete null holds (i.e., $\widehat{\pi}_1 = 0$), we expect $c_q \approx q$ to ensure the true FDR satisfies $\text{FDR}(\boldsymbol{\alpha}, c_q) \leq q$. When the proportion of true alternatives increases (i.e., $\widehat{\pi}_1$ becomes larger), the scaling factor c_q can be adjusted accordingly, allowing for a more permissive p-value threshold. In the extreme case where $\pi_1 = 1$ (i.e., all hypotheses are true alternatives), c_q can theoretically grow arbitrarily large, as no false positives exist in this scenario.

Our objective is to determine the optimal value of c_q, thereby providing an accurate estimate of the p-value threshold that balances FDR control and statistical power. Although the true test statistics $\boldsymbol{\alpha}$ are not directly observable, we have access to their estimates $\widehat{\boldsymbol{\alpha}} = \boldsymbol{\alpha} + \mathbf{u}$,

[4] When the p-values are sorted in ascending order ($p_1 \leq p_2 \leq \cdots \leq p_N$), the CDF at $p = p_i$ counts exactly i values such that $p_j \leq p_i$ for $j \leq i$. Therefore $\widehat{F}_{\widehat{\boldsymbol{\alpha}}}(p_i) = \frac{i}{N}$.

where $\mathbf{u} \sim \mathcal{U}$. We employ a cross-sectional resampling approach, which allows us to generate bootstrap samples $\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \mathbf{u}^{(3)}, \ldots$ from the null distribution $\mathcal{U}$.

Cross-sectional resampling. Consider the estimated residuals $\widehat{\varepsilon}_{i,t} = y_{i,t} - \widehat{\boldsymbol{\beta}}^\top \mathbf{x}_t$ from Eq.(1), where $\widehat{\boldsymbol{\beta}}^\top = \left(\widehat{a}_i, \widehat{b}_i^\top\right) \in \mathbb{R}^{p+1}$ are the ordinary least squares (OLS) estimates of the coefficients. We consider a conventional cross-sectional bootstrap (e.g., Kosowski et al., 2006; Fama and French, 2010; Barras et al., 2010). For each simulation $b = 1, \ldots, B$, we sample (with replacement) T time steps $\mathcal{T}_b = (t_1^{(b)}, t_2^{(b)}, \ldots, t_T^{(b)}) \in [T] \times [T] \times \cdots \times [T]$ to compute a variation of residuals:

$$\widehat{e}_i^{(b)} = \frac{1}{T} \sum_{t \in \mathcal{T}_b} \widehat{\varepsilon}_{i,t}, \qquad b = 1, \ldots, B. \tag{8}$$

Although the original motivation in Fama and French (2010) was to use this method for percentile inference, we show that the cross-sectional correlation of $\widehat{e}_i^{(b)}$ provides a consistent estimate of the dependence structure between portfolio returns.

Theorem 1. (Consistency of residual bootstrap) *Assume the following conditions hold:*

(i) *The matrix* $\boldsymbol{X} = \left(\mathbf{x}_1^\top; \ldots; \mathbf{x}_T^\top\right) \in \mathbb{R}^{T \times p}$ *has rank* p *and* $\|\boldsymbol{X}\|$ *is bounded, where* $\|\cdot\|$ *denotes the spectral norm.*

(ii) $\mathbb{E}[\boldsymbol{X}^\top \varepsilon_i] = 0$ *for all* $i \in [N]$*, where* $(\varepsilon_{i,t})_{t \in [T]}$ *are the residual errors.*

(iii) *The residual errors* $(\varepsilon_{i,t})_{t \in [T]}$ *are independent and identically distributed as* $\mathcal{N}(0, \sigma_i^2)$*.*

Then, for each of the $b = 1, 2, \ldots, B$ *bootstrap residuals* $\widehat{e}_i^{(b)}$*, the following properties hold:*

1. *The bootstrap mean satisfies:*

$$\mathbb{E}_{\text{boot}}\left[\widehat{e}_i^{(b)}\right] = \frac{1}{|\mathcal{S}(T)|} \sum_{\mathcal{T}_b \in \mathcal{S}(T)} \widehat{e}_i^{(b)} = 0, \tag{9}$$

where $\mathbb{E}_{\text{boot}}[\cdot]$ *denotes the expectation over the randomness of bootstrap resampling, and* $\mathcal{S}(T) = [T]^T$ *is the set of all possible bootstrap samples of size* T *(with* $|\mathcal{S}(T)| = T^T$*).*

2. The bootstrap covariance satisfies, with probability at least $1-\delta$*:*

$$T\,\mathrm{Cov}_{\mathrm{boot}}(\widehat{e}_i^{(b)},\widehat{e}_j^{(b)}) = \mathrm{Cov}(\varepsilon_i,\varepsilon_j) + O\left(\frac{\log(1/\delta)}{\sqrt{T}}\right), \tag{10}$$

where $\mathrm{Cov}_{\mathrm{boot}}(\cdot,\cdot)$ *denotes the empirical covariance over bootstrap samples.*

The first and the second assumptions in Theorem 1 are general. The third assumption states that the serial correlation is weak, whereas the correlation among portfolios at the same period can be large. Appendix B provides the proof.

FDR and optimal threshold. Theorem 1 demonstrates that each bootstrap sample $\widehat{e}_i^{(b)}$, $b = 1,\ldots,B$, preserves the correlation structure present in the data. We leverage this result to generate correlated perturbed test statistics $\widehat{\boldsymbol{\alpha}}^{(b)} = \widehat{\boldsymbol{\alpha}} - \boldsymbol{u}^{(b)}$, where $\boldsymbol{u}^{(b)} = \left(u_1^{(b)}, u_2^{(b)}, \ldots, u_N^{(b)}\right)$ has ith element $u_i^{(b)} = \frac{\widehat{e}_i^{(b)}}{\widehat{\sigma}_i^{(b)}/\sqrt{T}}$ and $\widehat{\boldsymbol{\alpha}}$ is the vector of observed studentized test statistics (see Eq.(2)). We use these bootstrap samples to estimate $\widehat{N_0}$, which is then used for estimating the optimal threshold c_q^*.

To ensure the number of null hypotheses $|\mathcal{H}_0(\widehat{\boldsymbol{\alpha}}^{(b)})|$ is at least as large as the true number of null hypotheses $N_0 = |\mathcal{H}_0(\boldsymbol{\alpha})|$, any test statistic $\widehat{\alpha}_i^{(b)}$ is set to 0 if its magnitude does not exceed the bootstrap noise level, i.e., $|\widehat{\alpha}_i| \le |u_i^{(b)}|$. This ensures that weak signals are treated as null hypotheses in the bootstrap sample. Algorithm 1 describes the sampling procedure.

Algorithm 1: Sampling $\widehat{\boldsymbol{\alpha}}^{(b)}, b = 1,\ldots,B$

Input : $\widehat{\boldsymbol{\alpha}}$, $q \in (0,1)$, $B \in \mathbb{N}^+$

Sample $\boldsymbol{u}^{(b)} \sim \mathcal{U}$.

Compute $\widehat{\boldsymbol{\alpha}}^{(b)} \leftarrow \widehat{\boldsymbol{\alpha}} - \boldsymbol{u}^{(b)}$.

Set $\widehat{\alpha}_i^{(b)} \leftarrow 0$ for all $\{i \in [N] : |\widehat{\alpha}_i| \le |u_i^{(b)}|\}$. `/* Two-sided testing */`

Output: $\widehat{\boldsymbol{\alpha}}^{(b)}$ with adjusted null hypothesis structure.

Algorithm 1 ensures that by setting $\widehat{\alpha}_i^{(b)} = 0$ for weak signals ($|\widehat{\alpha}_i| \le |u_i^{(b)}|$), the algorithm guarantees that the number of null hypotheses $|\mathcal{H}_0(\widehat{\boldsymbol{\alpha}}^{(b)})|$ in the bootstrap sample is at least as large as $N_0 = |\mathcal{H}_0(\boldsymbol{\alpha})|$, the true number of null hypotheses. The use of $\boldsymbol{u}^{(b)} \sim \mathcal{U}$ ensures that

the bootstrap samples mimic the dependence structure of the original data, as $\mathcal{U}$ is designed to reflect the variability and correlation in the observed errors.

As highlighted by Genovese (2004), the main challenge of FDR control under dependence lies in estimating N_0, the number of true null hypotheses. Lemma 2 provides theoretical guarantees that, with probability arbitrarily close to one, at least one bootstrap sample $\widehat{\boldsymbol{\alpha}}^{(b)}$, $b = 1, \ldots, B$, will contain at least $N_0/2$ null hypotheses.

Lemma 2. (Number of Null Hypotheses) *With probability at least* $1 - \frac{2}{B+1}$, *there exists at least one bootstrap simulation* $b \in [B]$ *such that* $|\mathcal{H}_0(\widehat{\boldsymbol{\alpha}}^{(b)})| \geq N_0/2$.

Appendix B provides the complete proof and introduces a "dueling estimator," which we use to derive a lower bound for the number of null hypotheses such that $|\mathcal{H}_0(\hat{\boldsymbol{\alpha}}^{(b)})| \geq N_0/2$ for some $b \in [B]$. After generating the bootstrap samples $\widehat{\boldsymbol{\alpha}}^{(b)}$, $b = 1, \ldots, B$, the estimated FDR is obtained by replacing the expectation with its empirical counterpart using W bootstrapped samples. Algorithm 2 outlines the corresponding procedure.

We can now determine the p-value threshold that controls the FDR at the desired level q. This is achieved by sampling $\widehat{\boldsymbol{\alpha}}^{(b)}$, $b = 1, \ldots, B$ (see Algorithm 1) and then estimating $\widehat{\mathrm{FDR}}(\widehat{\boldsymbol{\alpha}}^{(b)}, c_q)$ based on W draws of $\widehat{\boldsymbol{\alpha}}^{(w)} = \widehat{\boldsymbol{\alpha}}^{(b)} + \boldsymbol{u}^{(w)}$ (see Algorithm 2). This procedure ensures that the estimated FDR accounts for both the variability in $\widehat{\boldsymbol{\alpha}}^{(b)}$ and the dependence structure modeled by $\mathcal{U}$. Algorithm 3 outlines the complete dueling double-bootstrap (DDB) algorithm, which is used to determine the optimal p-value threshold. Theorem 3 provides theoretical guarantees on the accuracy of the FDR estimates. Appendix B provides the proof.

Theorem 3. (Main Theorem) *For any* $\boldsymbol{\alpha}'$, *assume that* $\mathrm{FDR}(\boldsymbol{\alpha}', c)$ *is non-decreasing in* c. *Let*

$$c_q^* = \sup_{c_q} \{c_q : \mathrm{FDR}(\boldsymbol{\alpha}, c_q) \leq q\}$$

be the desired threshold for the true FDR at level q. *If there exists* $\widehat{\boldsymbol{\alpha}}^{(b)}$, $b \in [B]$, *such that*

$$\mathrm{FDR}(\widehat{\boldsymbol{\alpha}}^{(b)}, c_{q'}^*) \geq \frac{q'}{2}, \tag{11}$$

Algorithm 2: Calculate $\widehat{\mathrm{FDR}}(\widehat{\boldsymbol{\alpha}}^{(b)}, c_q)$ for $b \in [B]$

Input : $c_q \in (0,1)$, $\widehat{\boldsymbol{\alpha}}^{(b)}$, $W \in \mathbb{N}^+$.

for $w = 1, \ldots, W$ **do**

(i) Sample $\boldsymbol{u}^{(w)} \sim \mathcal{U}$.

(ii) Compute perturbed statistics: $\widehat{\boldsymbol{\alpha}}^{(w)} = \widehat{\boldsymbol{\alpha}}^{(b)} + \boldsymbol{u}^{(w)}$.

(iii) Calculate the p-value threshold:

$$p \leftarrow \sup\{p' : \widehat{F}_{\widehat{\boldsymbol{\alpha}}^{(w)}}(p') \geq p'/c_q\},$$

where $\widehat{F}_{\widehat{\boldsymbol{\alpha}}^{(w)}}(p')$ is the empirical CDF of p-values based on $\widehat{\boldsymbol{\alpha}}^{(w)}$.

(iv) Calculate the number of discoveries:

$$R \leftarrow |\{i \in [N] : p_i(\widehat{\alpha}_i^{(w)}) \leq p\}|,$$

and the number of false positives:

$$N_{1|0} \leftarrow |\{i \in [N] : p_i(\widehat{\alpha}_i^{(w)}) \leq p, i \in \mathcal{H}_0(\widehat{\boldsymbol{\alpha}}^{(b)})\}|.$$

(v) Compute the FDR for iteration w:

$$\widehat{\mathrm{FDR}}^{(w)} \leftarrow \frac{N_{1|0}}{\max(1, R)}.$$

Output:

$$\widehat{\mathrm{FDR}}(\widehat{\boldsymbol{\alpha}}^{(b)}, c_q) \leftarrow \frac{1}{W} \sum_{w=1}^{W} \widehat{\mathrm{FDR}}^{(w)}.$$

Algorithm 3: Dueling Double Bootstrap (DDB)

Input : $\widehat{\boldsymbol{\alpha}}$, $q \in (0,1)$, $B, W \in \mathbb{N}^+$, $c_q \in \{0, 0.01, \ldots, 0.99, 1\}$.

for $b = 1, \ldots, B$ **do**

(i) Sample $\widehat{\boldsymbol{\alpha}}^{(b)}$ using Algorithm 1.

(ii) Determine $c_q^{(b)} \leftarrow \sup\{c_q \in (0,1) : \widehat{\mathrm{FDR}}(\widehat{\boldsymbol{\alpha}}^{(b)}, c_q) \leq q/2\}$ using Algorithm 2.

Output: Final p-value threshold p using Eq. (7) with $c_q \leftarrow \min_{b=1}^{B} c_q^{(b)}$.

for all $q' \in (0,1)$*, then with probability at least* $1-2\delta$*, our proposed method controls the FDR at most:*

$$q + 2\sqrt{\frac{\log((BS)/\delta)}{2W}}, \tag{12}$$

where B *is the number of bootstrap samples for* $\widehat{\boldsymbol{\alpha}}^{(b)}$*,* S *is the number of repetitions in the binary search, and* W *is the number of bootstrap samples used to estimate* FDR*.*

Eq. (11) suggests that, if Lemma 2 holds, then at least one bootstrap sample $\hat{\boldsymbol{\alpha}}^{(b)}$ gives a lower bound of FDR. As a result, Eq. (12) guarantees the FDR control at q when W is large.

We note that Algorithm 3 employs a double nested bootstrap approach. Specifically, it samples $\widehat{\boldsymbol{\alpha}}^{(b)}$ B times, and for each $\widehat{\boldsymbol{\alpha}}^{(b)}$, it estimates the FDR using W bootstrap samples. Consequently, the computational complexity of the algorithm is $O(BW)$. The parameters B and W serve distinct purposes. B ensures a lower bound on N_0, the number of true null hypotheses. By Lemma 2, $B = O(1/q)$ samples suffice to avoid underestimating N_0 with probability at least $1-q$. W controls the accuracy of the FDR estimation. Theorem 3 implies that $W = \widetilde{O}(1/q^2)$ samples – where $\widetilde{O}$ suppresses polylogarithmic factors – are sufficient to bound the FDR estimation error to $O(q)$.

Combining these results, the complexity of Algorithm 3 is $O(BW) = \widetilde{O}(1/q^3)$. Each FDR calculation requires $\widetilde{O}(N)$ operations due to the step-up procedure on the sorted p-values, where N is the number of hypotheses. Thus, the total computational complexity is $\widetilde{O}(N/q^3)$.

3 Simulation study

To complement the asymptotic guarantees of Theorem 3, we examine the finite-sample properties of our proposed method through a comprehensive simulation study. Following Romano et al. (2008), we simulate portfolio returns with sample sizes T drawn from a multivariate normal distribution. The returns have a mean vector $\boldsymbol{a} = (a_1, \ldots, a_N)$ and a correlation matrix $\Sigma = (\sigma_{i,j}) \in \mathbb{R}^{N \times N}$, where $N = 50$. The diagonal elements of Σ are fixed at $\sigma_{i,i} = 1$, while the off-diagonal elements $\sigma_{i,j}$ are set to a common correlation ρ, with $\rho \in \{0, 0.5, 0.9\}$, representing

independent, moderately correlated, and highly correlated hypotheses, respectively. For each alternative hypothesis $i \in \mathcal{H}_1$, the true mean a_i is independently sampled from $2 \times \text{Unif}(0, 1)$.

Simulation results are averaged over 2,000 independent runs, and we consider three scenarios for the proportion of true null hypotheses, $\pi_0 = N_0/N$, with $\pi_0 \in \{0.25, 0.5, 1\}$. Our proposed `DDB` method is compared against a range of baseline and widely used testing approaches:

(i) *Naive single hypothesis testing*: A fixed threshold of $p = 0.05$ (`Single`).

(ii) *Linear step-up methods*: Benjamini and Hochberg (1995) (`BH`), Benjamini and Yekutieli (2001) (`BY`), and Benjamini et al. (2006) (`BKY`).

(iii) *Storey methods*: Storey (2002) (`Storey`) with a fixed reference point $p_0 = 0.5$, and Storey et al. (2004) (`STS`), which adaptively selects p_0 using bootstrapped p-values.[5]

(iv) *Bootstrap methods*: Yekutieli and Benjamini (1999) (`YB`), calibrated using a pointwise $1 - q/2$ upper percentile for the FDR and confidence at $q/2$.

Additionally, we evaluate an aggressive variant of our method, labeled `DDBA`, which relaxes the threshold from $q/2$ to q. For our `DDB` and `DDBA` methods, we set $B = 20$ and $W = 500$, ensuring $B = 1/q$ and $W \geq 1/q^2$.

Table 1 presents the results for $T = 100$. The performance criteria include: (1) the p-value threshold, (2) the empirical FDR compared to the nominal level $q = 0.05$, and (3) the empirical statistical power, measured as the average number of false hypotheses rejected. Several noteworthy patterns emerge.

Traditional methods such as `BH`, `BY`, and `BKY` are highly conservative, exhibiting the strictest FDR control across all scenarios. However, this conservatism comes at the expense of statistical power, especially under high correlation. For example, with $\rho = 0.9$ and $\pi_0 = 0.25$, `BY` rejects only 2.5% of false hypotheses, compared to 9.2% for `DDB` and 13.3% for `DDBA`.

In contrast, methods like `Storey` and `STS` achieve the highest statistical power but fail to adequately control FDR. For instance, under $\pi_0 = 0.25$ and $\rho = 0.9$, `STS` rejects 32.9% of

[5]Both `BKY` and `STS` use 500 bootstrap samples in the simulations.

Panel A: True null hypotheses ratio $\pi_0 = 0.25$

	$\rho = 0$			$\rho = 0.5$			$\rho = 0.9$		
Method	Thr(%)	FDR(%)	Rej(%)	Thr(%)	FDR(%)	Rej(%)	Thr(%)	FDR(%)	Rej(%)
`Single`	4.09	7.04	16.64	3.86	10.62	16.28	2.92	6.51	16.44
`BH`	0.09	1.55	2.68	0.21	0.99	4.88	0.30	0.83	7.04
`BKY`	0.09	1.55	2.62	0.41	1.22	5.78	0.74	1.27	8.56
`BY`	0.00	0.43	0.66	0.01	0.07	1.44	0.02	0.17	2.54
`Storey`	0.22	2.19	4.08	3.31	2.51	11.44	7.05	4.38	20.34
`STS`	15.18	6.38	20.78	20.66	6.82	27.44	19.64	7.86	32.88
`YB`	0.02	1.06	1.38	0.06	0.68	2.96	0.25	0.91	6.76
`DDB`	0.09	1.47	2.70	0.42	1.05	5.82	0.81	1.33	9.18
`DDBA`	0.45	2.86	5.78	1.25	2.02	9.46	1.84	2.12	13.32

Panel B: True null hypotheses ratio $\pi_0 = 0.5$

	$\rho = 0$			$\rho = 0.5$			$\rho = 0.9$		
Method	Thr(%)	FDR(%)	Rej(%)	Thr(%)	FDR(%)	Rej(%)	Thr(%)	FDR(%)	Rej(%)
`Single`	3.85	19.22	12.58	3.63	20.40	12.80	2.73	9.71	12.30
`BH`	0.05	2.60	1.58	0.12	2.43	2.96	0.18	1.38	4.10
`BKY`	0.05	2.56	1.56	0.21	2.70	3.34	0.41	1.79	5.02
`BY`	0.00	0.32	0.42	0.01	0.63	0.90	0.01	0.32	1.02
`Storey`	0.08	3.44	2.02	1.90	4.75	7.02	6.45	8.66	17.24
`STS`	10.93	10.06	14.08	15.91	11.50	20.82	19.24	15.44	30.60
`YB`	0.01	1.64	0.86	0.04	1.83	1.84	0.19	1.85	4.24
`DDB`	0.04	2.32	1.42	0.22	2.46	3.28	0.52	2.20	5.56
`DDBA`	0.18	4.99	2.90	0.69	4.40	5.70	1.23	3.68	8.82

Panel C: True null hypotheses ratio $\pi_0 = 1$

	$\rho = 0$			$\rho = 0.5$			$\rho = 0.9$		
Method	Thr(%)	FDR(%)	Rej(%)	Thr(%)	FDR(%)	Rej(%)	Thr(%)	FDR(%)	Rej(%)
`Single`	3.11	90.95	4.98	2.37	63.65	5.12	0.83	18.71	4.30
`BH`	0.00	5.05	0.12	0.01	3.45	0.36	0.06	1.65	1.38
`BKY`	0.00	5.00	0.12	0.03	3.45	0.44	0.12	1.60	1.56
`BY`	0.00	1.05	0.02	0.00	0.80	0.06	0.00	0.40	0.34
`Storey`	0.00	5.55	0.12	1.10	8.50	3.22	6.19	17.15	17.08
`STS`	7.23	14.65	7.92	10.85	16.95	13.38	7.47	18.60	18.60
`YB`	0.00	3.45	0.08	0.00	2.90	0.16	0.03	2.10	0.92
`DDB`	0.00	3.90	0.08	0.06	3.45	0.64	0.29	3.35	2.98
`DDBA`	0.01	8.00	0.18	0.19	6.45	1.40	0.61	4.90	4.54

Table 1: **Simulation results**. The table reports the simulation results for $T = 100$. The performance of each method is reported in terms of p-value threshold (Thr-p), proportion or rejected hypotheses (Rej) and estimated FDR. These are average values out of 2,000 simulation runs. The table reports the results for different level of hypotheses correlation ($\rho = 0, 0.5, 0.9$).

false hypotheses, far exceeding the 13.3% achieved by `DDBA`. However, this is coupled with an FDR of 7.86%, significantly above the target level. Lenient methods like `Storey` and `STS`

generally exhibit overly aggressive rejection rates, particularly under high correlation or when the proportion of true nulls (π_0) increases, the estimator $\widehat{\pi}_1$ highly overestimates the number of true alternative hypotheses π_1, leading to poor FDR control.

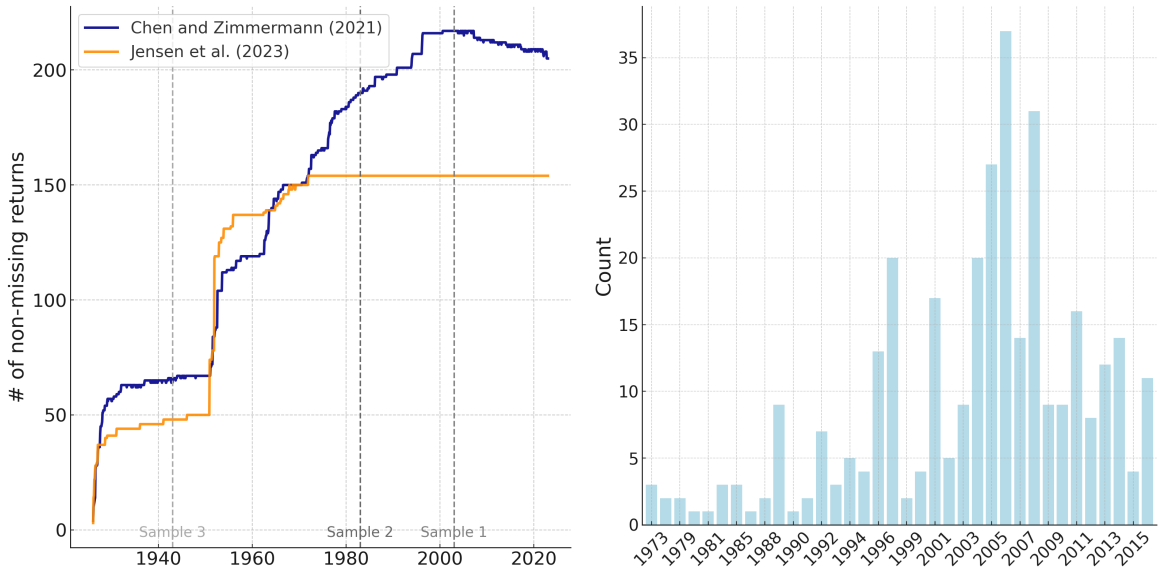
In comparison, `DDB` consistently maintains FDR close to the target level across all scenarios. `DDBA` offers slightly higher power while still achieving reasonable FDR control. For instance, under $\pi_0 = 0.5$ and $\rho = 0.9$, `DDBA` achieves an FDR of 3.68% – notably lower than the lenient methods – while rejecting a higher proportion of false hypotheses. Both `DDB` and `DDBA` adapt well to varying levels of correlation and π_0, demonstrating relatively stable FDR and rejection rates. For example, under $\pi_0 = 0.5$ and $\rho = 0.5$, `DDB` achieves an FDR of 2.46% while rejecting 3.28% of false hypotheses, maintaining robust control even as π_0 increases.

Results for shorter time series ($T = 50$) are presented in Table C.1 in Appendix C.1. These findings largely confirm the trends observed in Table 1. Our `DDB` and `DDBA` methods strike a robust balance between FDR control and statistical power. While `DDB` is the safer choice for strict FDR control, `DDBA` provides an effective compromise between control and rejection rates, outperforming both restrictive and lenient methods across most scenarios.

4 Revisiting anomalies replicability

In this Section, we investigate whether correlated portfolio anomaly returns withstand the multiple hypothesis testing standards imposed by our `DDB` approach. To this end, we leverage three alternative datasets of characteristic-managed portfolios. The first, compiled by Jensen et al. (2023) (JKP henceforth), includes 153 U.S. stock market portfolios constructed using various weighting schemes. Detailed descriptions of these portfolios can be found at `http://jkpfactors.com`. The second dataset, available at `http://openassetpricing.com`, retains the original research designs of anomalies proposed in the literature (see Chen and Zimmermann, 2021, for more details) (CZ henceforth).[6] The third, available at `http://openbondassetpricing.com`, includes the excess returns on 15 tradable corporate bond factors

[6] We use version 1.3.0, released in August 2023.



(a) Size of the cross-section (b) Year of publication in `openassetpricing.com`

Figure 1: **Sample size and publication years**. The left panel reports the number of non-missing portfolio returns for characteristic-managed portfolios collected in the datasets by Jensen et al. (2023) (dark orange) and Chen and Zimmermann (2021) (dark blue). The vertical dashed lines represent the alternative sample periods considered in the empirical analysis. The right panel reports the number of published anomalies per year as reported by `openassetpricing.com`.

(see Dickerson et al., 2023a, for more details).

A snapshot of portfolio anomalies. Figure 1 illustrates the temporal composition of the cross-section of equity portfolios (left panel) and the publication timeline of the anomalies as from Chen and Zimmermann (2021). Our main empirical analysis focuses on the period from January 1984 to December 2022 (Sample 2, marked by the vertical dashed line in the left panel). This period balances a sufficiently broad cross-section with a moderately long time series. Notably, this period encompasses the majority of newly published risk factors and anomalies, as shown in the right panel of Figure 1. The panel of bond portfolios is balanced and spans the period from January 1986 to December 2022.

We note that the characteristic-managed portfolios reported by CZ are consistent with the original studies, which predominantly employ equal-weighted portfolios. However, Hou et al. (2020) demonstrate that equal-weight portfolios tend to enhance replicability, with raw portfolio returns exhibiting greater statistical significance compared to those of value-weight

portfolios. Conversely, pure value-weighting can lead to excessively concentrated portfolios, potentially obscuring the contribution of specific characteristics to the cross-sectional variation in stock returns. For this reason, in addition to the original CZ portfolios, we follow Jensen et al. (2023) and also test for anomaly replicability using capped value-weight, equal-weight, and pure value-weight portfolio constructions for the 153 anomalies reported by JKP.

For corporate bond returns, portfolios are constructed based on the total bond portfolio return minus the one-month risk-free rate of return. We also follow van Binsbergen et al. (2025) and repeat our analysis with duration-adjusted returns, where the bond excess return is computed as the total bond return minus the return on a portfolio of duration-matched U.S. Treasury bonds. For more details on bond portfolio construction, we refer the interested reader to the description in Dickerson et al. (2023a).

Different portfolio construction methods can significantly influence the correlation structure of anomaly returns, with important implications for statistical inference. Appendix A illustrates this point. Figure A.8 presents heatmaps of the cross-sectional correlation of the 153 portfolio anomalies from JKP (left panel) and the 176 portfolio anomalies from CZ (right panel) used in Section 4.1. The evidence reveals strikingly heterogeneous correlations among portfolio anomalies. Although modest on average, there are distinct clusters of high correlations within anomalies sharing similar underlying signals, such as momentum- and volatility-based strategies. Notably, these correlation clusters suggest that what appears to be multiple distinct anomalies may in fact capture common underlying risk factors or market inefficiencies. This echoes the evidence in Jensen et al. (2023), who argue that portfolio anomalies can be aggregated according to common "themes". We will discuss this in more detail in Section **??**.

Turning to fixed income markets, the returns on bond factors exhibit an even more pronounced correlation structure. Figure A.9 presents the pairwise return correlations of the tradeable bond portfolio returns from Dickerson et al. (2023a). The left panel displays excess portfolio returns, whereas the right panel shows duration-adjusted excess returns (see van Binsbergen et al., 2025). Value, carry, credit risk, default risk, and downside risk factors are

highly correlated with each other, with these correlations intensifying for duration-adjusted returns.

Economically, the high correlation observed within subsets of equity and bond portfolios challenges the conventional view that these represent truly independent risk premiums. Statistically, correlated hypotheses dilute the effectiveness of individual test statistics and increase the likelihood that multiple rejections reflect shared signals rather than distinct discoveries (e.g., Bai and Ng, 2010; Harvey et al., 2016). These preliminary findings underscore the critical importance of accounting for correlation structures when evaluating the statistical significance of apparent anomalies in both equity and fixed income markets.

4.1 Replication rates

Jensen et al. (2023) anomaly portfolios. We first test for anomaly replicability based on equal-weight, value-weight, and capped value-weight construction schemes for 153 anomalies for the U.S. stock market as in JKP. The analysis is conducted using raw portfolio returns and CAPM alphas, where market risk is measured by the returns on the market portfolio in excess of the risk-free rate.[7]

Figure 2 displays the monthly average returns of each portfolio anomaly, sorted in ascending order. Anomalies successfully replicated by our dueling double-bootstrap (`DDB`) approach are highlighted in dark orange, while non-replicated anomalies are shown in dark blue. A portfolio anomaly is classified as replicated, i.e., a true discovery, if its average raw or risk-adjusted return is statistically significant under FDR control. For clarity, the y-axis in the subplots displays every other characteristic label.

The results reveal significant differences across weighting schemes. Value-weight portfolios exhibit the lowest replication rate for raw returns (28.7%), reflecting the concentrated influence of large-cap stocks. Equal-weight portfolios achieve the highest replication rate (75.16%),

[7]The returns on the market portfolio and the 1-month risk-free rate are from Kenneth French's website https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

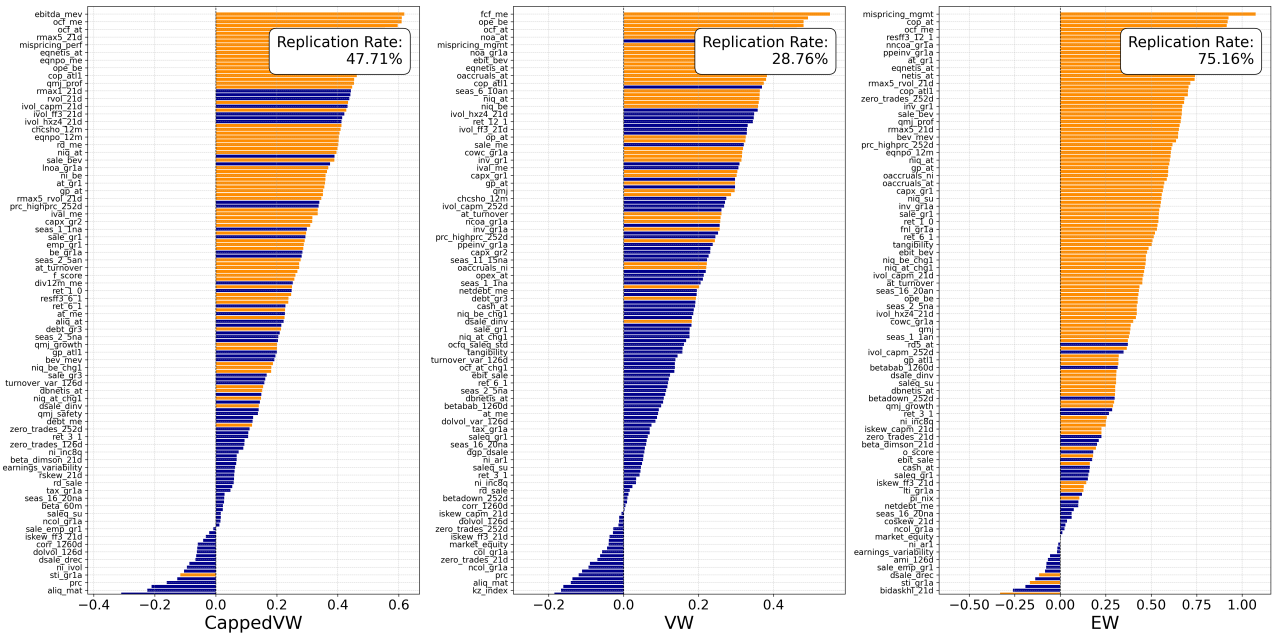


Figure 2: **Replication rates for raw returns**. This figure reports the monthly raw returns (in %) for each portfolio anomaly, with replication rates based on statistical significance under our DDB approach. Anomalies successfully replicated are highlighted in dark yellow. The sample spans January 1984 to December 2022. The left panel shows results for capped value-weight portfolios, the middle panel for pure value-weight portfolios, and the right panel for equal-weight portfolios. For clarity, the y-axis in the subplots displays every other characteristic label.

driven by their greater sensitivity to smaller stocks, albeit at the cost of limited practical tradeability (e.g., Avramov et al., 2023). Capped value-weight portfolios, which balance exposure between small- and large-cap stocks, yield a replication rate of 47.7%, representing a middle ground between the two extremes. These findings align with Hou et al. (2020), who report lower replication rates when departing from equal-weighted portfolios.

Economic theory suggests that risk-adjusted returns are better suited for detecting anomalies (Jensen et al., 2023). Figure 3 reports the replication rates when raw returns are replaced by monthly CAPM alphas. The replication rate for capped value-weight portfolios shows a substantial improvement compared to raw returns (75.16% vs. 47.71%). Similarly, the replication rates for value-weight (67.97%) and equal-weight (86.93%) portfolios also increase significantly when using CAPM alphas. These findings underscore that, for many of the non-replicated anomalies under raw returns, the true underlying signal was buried under the noise introduced by exposure to market risk.

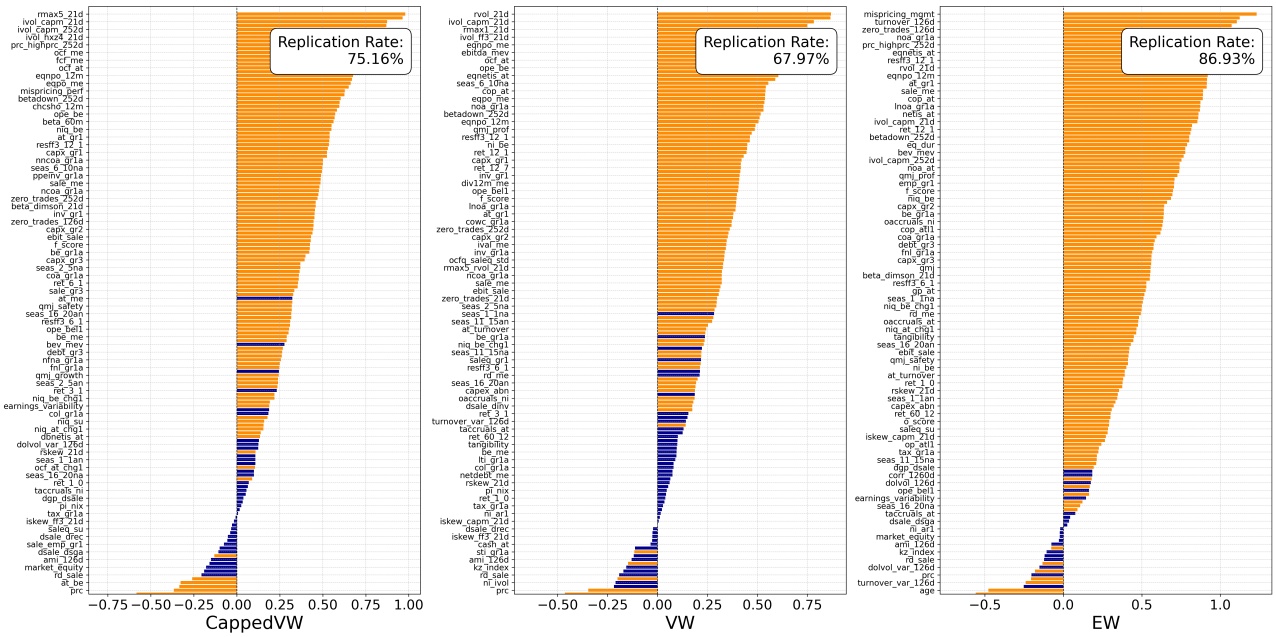


Figure 3: **Replication rates for CAPM alphas**. This figure reports the monthly CAPM alphas (in %) for each portfolio anomaly. The anomalies that are successfully replicated according to our DDB approach are highlighted in dark yellow. The sample size is from January 1984 to December 2022. The left panel shows the results for the capped value-weight portfolios, the middle panel for the value-weight portfolios, and the right panel for the equal-weight portfolios. For clarity, the y-axis in the subplots displays every other characteristic label.

More generally, one of the leading criticisms of factor research replicability is that results are driven by illiquid small stocks whose behavior largely reflects market frictions and microstructure rather than economic fundamentals or investor preferences (e.g., Green et al., 2017; Amihud, 2002). In particular, Hou et al. (2020) argue that they find a low replication rate because they limit the influence of micro-caps. Our results show that anomalies demonstrate a high replication rate even when the impact of small caps is mitigated via value weighting.

Appendix D show the cumulative distribution functions (CDFs) of the test statistics under the null hypothesis, generated through our dueling double-bootstrap approach. Figure D.1 compares the null distributions to the empirical t-statistic CDF for raw returns across various portfolio constructions (capped value-weight, pure value-weight, and equal-weight). The empirical alpha CDF (orange) is consistently shifted to the right relative to the null distribution CDF (blue).

The rightward shift of the empirical t-statistic CDF implies that portfolio returns are systematically larger than what would be expected under the null hypothesis of no anomaly, supporting the presence of significant anomalies in the data. This pattern is observed across all portfolio constructions, though the magnitude of the shift may vary, suggesting that certain portfolio weighting schemes, such as equal-weight, may amplify the anomalies' detectability.

Figure D.2 reports the null and empirical t-statistic distributions for the CAPM alphas. The separation between the empirical alpha (orange) and the null distribution (blue) CDF is noticeably broader. This broader range reflects stronger deviations from the null hypothesis, suggesting that portfolio anomalies can be more strongly replicated when returns are adjuted by their exposure to market risk.

Appendix D presents additional results based on monthly risk-adjusted returns that condition on size, value, and momentum factors, in addition to market risk (Carhart, 1997). Figure D.3 shows that the replication rates decrease slightly compared to those based on CAPM alphas but remain significantly higher than those for raw returns.

This marginal reduction in replication rates is likely due to some portfolio anomalies, such as those constructed around momentum signals, being mechanically related to the risk factors included in the four-factor model. For example, the momentum factor explicitly captures return variation linked to momentum anomalies, leaving less unexplained variation available to classify such portfolios as true anomalies. These findings highlight a trade-off in benchmark selection for anomaly replication: while the CAPM provides a simpler baseline for anomaly detection, multi-factor models like Carhart's four-factor model may overcorrect by attributing returns to factors that are closely tied to the anomalies themselves.

More broadly, these results align with prior findings that multi-factor models improve explanatory power but may reduce the detectability of anomalies in empirical studies (Green et al., 2017). They also emphasize the importance of selecting benchmarks that balance simplicity and robustness, as discussed in Harvey et al. (2016). Taken together, these findings support the argument that the CAPM may offer a more appropriate benchmark for anomaly

detection in cases where anomalies may overlap with risk factors in more extensive models.

Chen and Zimmermann (2021) characteristic-managed portfolios. Building on our analysis of anomaly replicability in the JKP dataset, we now turn to the 176 characteristic-managed portfolios from Chen and Zimmermann (2021), which have been continuously available from January 1984 to December 2022. Figure 4 reports the replication rates for raw returns and CAPM alphas. For clarity, the y-axis in the subplots displays every other characteristic label. The replication rate for raw portfolio returns is substantial at 82.4%, exceeding the 75.2% reported for equal-weight (EW) portfolios in Jensen et al. (2023). This suggests stronger evidence in favor of anomaly replicability for this dataset. The replication rate is even higher for CAPM alphas, reaching 87.5% under our `DDB` approach.

These findings reflect, in part, the larger cross-section of anomalies in the CZ dataset, which enhances statistical power by reducing sampling error, and, in part, the construction of characteristic-managed portfolios, which mostly include equal-weight schemes, which tend to produce higher replication rates due to their sensitivity to small-cap stocks (Hou et al., 2020).

Appendix D presents additional results for the CZ data based on monthly risk-adjusted returns that condition on size, value, and momentum factors, in addition to market risk (Carhart, 1997). Figure D.4 shows that, similar to the JKP portfolio anomalies, the replication rates conditioning on multiple factors are comparable to those based on CAPM alphas and are larger than those for raw returns.

Replicability by economic groups. Jensen et al. (2023) proposed a simple clustering approach to aggregate the large cross-section of anomalies and risk factors into a handful of so-called "themes." These clusters share a high degree of economic similarity and are named after the characteristics that mostly dominate the cluster: accruals, debt issuance, investment, leverage, low risk, momentum, profit growth, profitability, quality, seasonality, size, skewness, and value.

Grouping anomalies by economic themes provides a structured way to analyze replication

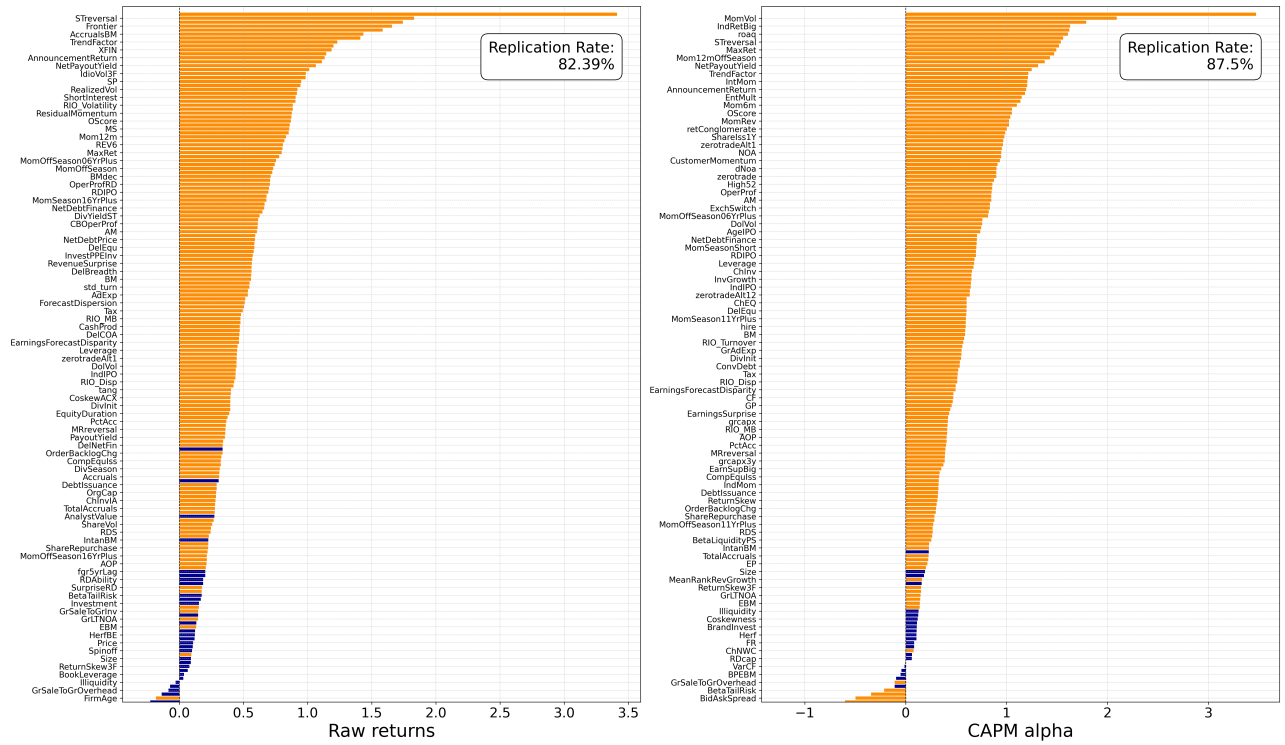


Figure 4: **Replication rates for Chen and Zimmermann (2021) dataset**. This figure reports the monthly portfolio average returns (left panel) and CAPM alphas (right panel) in % for each portfolio anomaly in the Chen and Zimmermann (2021) dataset. Anomalies successfully replicated according to our DDB approach are highlighted in dark yellow. The sample spans January 1984 to December 2022. For clarity, the y-axis in the subplots displays every other characteristic label.

rates, offering insights into whether certain economic mechanisms are inherently more robust or replicable across different portfolio constructions. In this section, we leverage this coarser aggregation of portfolio anomalies to investigate the extent to which replication rates differ across characteristics with homogeneous economic meaning. Figure 5 reports the replication rates clustered by themes for raw returns (light red) and CAPM alphas (light blue).

The replication rates vary significantly across themes and portfolio constructions, highlighting that certain anomaly characteristics may be more conducive to replication. For instance, when examining pure value-weight portfolios, momentum, profit growth, value, and low-risk anomalies exhibit higher replication rates when using CAPM alphas but relatively low replication rates when multiple hypothesis testing (MHT) is based on raw returns. This pattern is consistent for capped value-weight portfolios.

In contrast, the heterogeneity between anomaly themes diminishes when using equal-weight

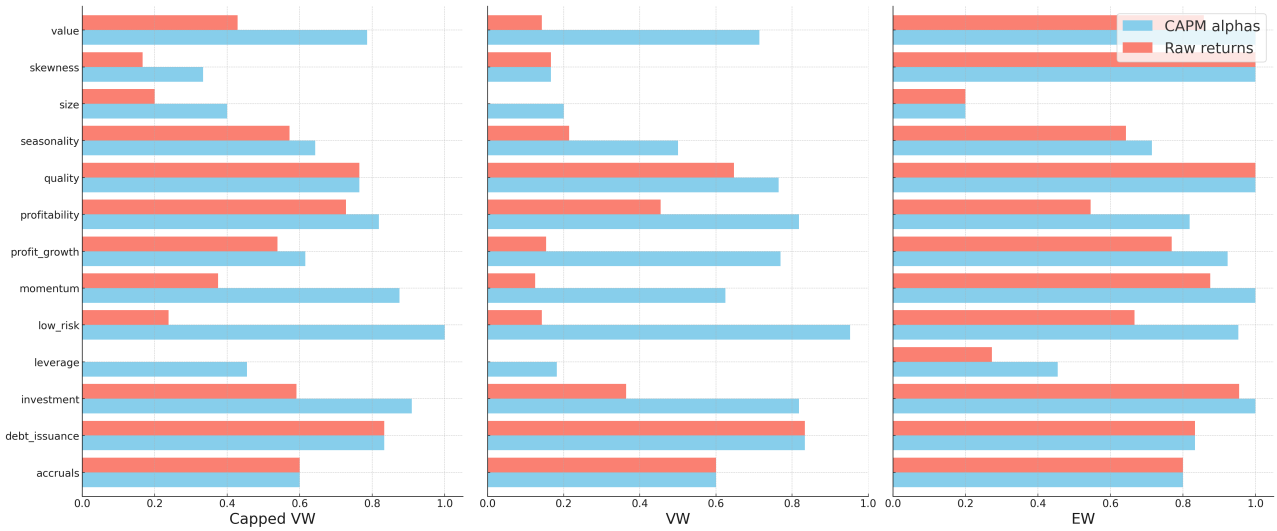


Figure 5: **Replication rates by anomaly theme**. This figure reports the replication rate (in decimals) for each anomaly theme as identified by Jensen et al. (2023). The figure shows replication rates for CAPM alphas (light blue) and raw returns (light red). The sample spans January 1984 to December 2022.

portfolios. This aligns with the broader empirical evidence showing that overweighting micro-caps tends to increase replication rates for raw returns relative to CAPM alphas (Hou et al., 2020; Green et al., 2017). The diminished heterogeneity across themes under equal-weight portfolios reflects their sensitivity to small-cap stocks, which often exhibit greater cross-sectional return variation, leading to higher replication rates. Interestingly, leverage anomalies become partly replicable under equal-weight schemes. However, the replicability of raw returns for size anomalies disappears for value-weight portfolios.

Similar to Jensen et al. (2023), leverage and size anomalies consistently exhibit low replication rates, regardless of whether raw returns or CAPM alphas are used, and irrespective of the portfolio construction scheme. When using capped value-weight and pure value-weight returns, none of the anomalies clustered in the leverage theme (e.g., cash-to-assets, net-debt-to-total-assets, and tangibility) achieve statistical significance at conventional thresholds. This consistently low replication rate suggests that leverage and size anomalies may either lack robust economic signals or rely on noisy firm-level accounting data.

To further investigate the replicability of anomalies across economic themes, in Appendix D.1 we analyze the replication rates by economic groups as indicated by Chen and Zimmermann

(2021). The poor replicability of size and leverage anomalies is corroborated by Figure D.6. The replicability of portfolio returns is remarkably similar between raw returns and CAPM alphas. This similarity is likely driven by the predominance of equal-weight portfolios in the dataset. Equal-weight portfolios inherently reduce the influence of market risk, leading to raw returns that are highly correlated with CAPM alphas.

Bond risk factors. We now turn our attention to a set of 15 tradable corporate bond risk factors. Similar to the equity anomaly replicability, we consider both raw returns and returns in excess of their exposure to a bond market portfolio. Note that with the exception of the term spread – calculated as the spread between the return on a long-term government bond and a short-term government bond – all bond risk factors considered are a function of bond characteristics such as credit ratings, duration, issuance size, past performances, or firm characteristics such as earnings announcements and equity momentum. For more details on bond factors, we refer the interested reader to the description in Dickerson et al. (2023a).

Portfolios are constructed based on the total bond portfolio return minus the one-month risk-free rate of return. We also follow van Binsbergen et al. (2025) and repeat our analysis with duration-adjusted returns, where the bond excess return is computed as the total bond return minus the return on a portfolio of duration-matched U.S. Treasury bonds.

The left panels of Figure 6 report the monthly average returns of each bond portfolios, sorted in ascending order. Anomalies successfully replicated by our dueling double-bootstrap (`DDB`) approach are highlighted in dark orange, while non-replicated anomalies are shown in dark blue. We note that a risk factor is classified as replicated, i.e., a true discovery, if its average raw or risk-adjusted return is statistically significant under FDR control.

Contrary to equity portfolio anomalies, the replication rate for raw bond portfolio returns is rather small. For instance, for duration-adjusted returns only 20% of the risk factors generate statistically significant returns when controlling for FDR. The replication rate increases to 46.7% when using standard bond portfolio excess returns. Nevertheless, replication rates are

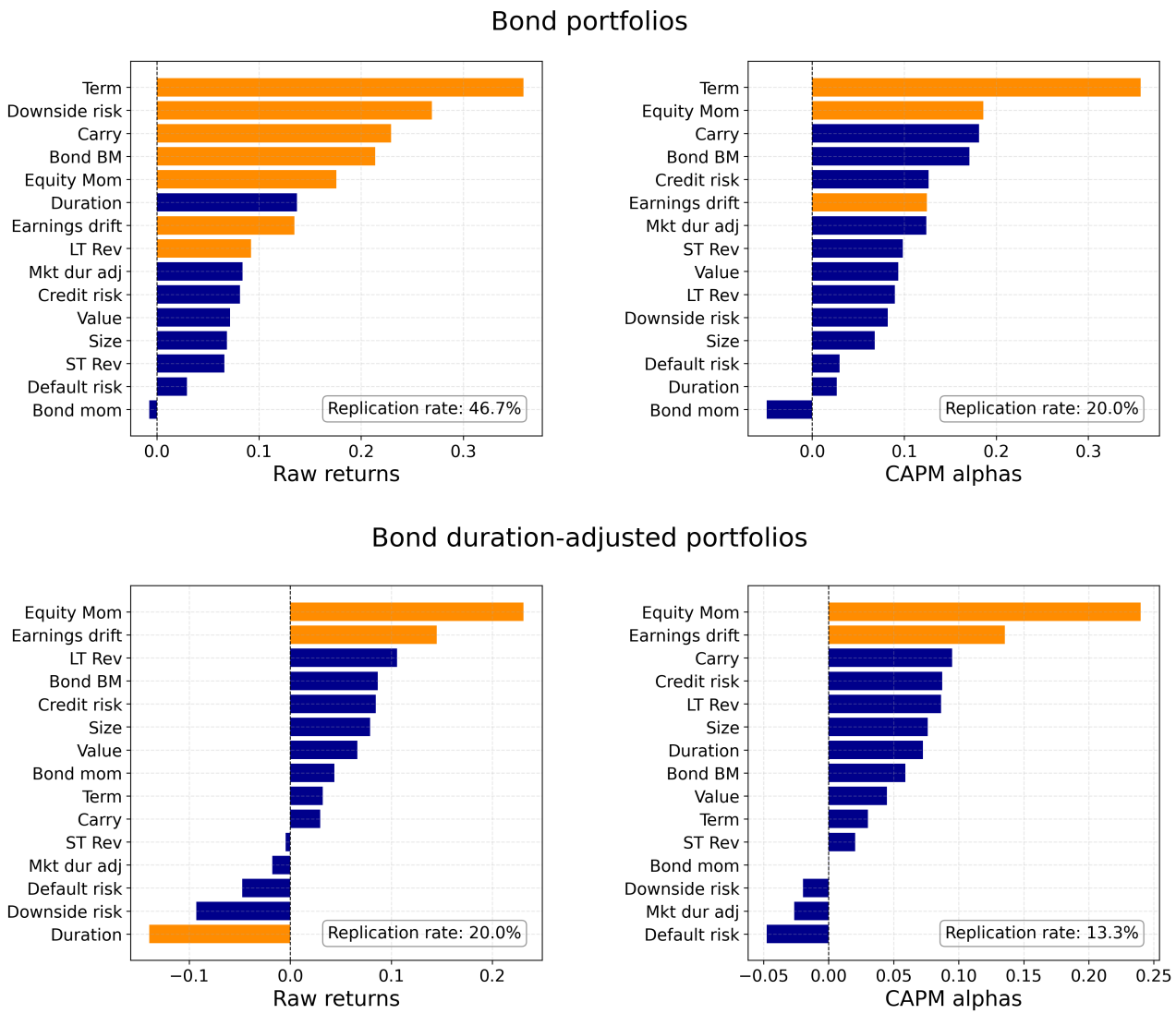


Figure 6: **Replication rates for bond portfolios**. This figure reports the monthly raw returns (in %) and the CAPM alphas for each bond portfolios, with replication rates based on statistical significance under our `DDB` approach. Risk factors successfully replicated are highlighted in dark yellow. The sample spans January 1986 to December 2022. The top panels show the results for standard portfolio returns, and the bottom panels for duration-adjusted portfolio returns.

substantially smaller compared to equity anomalies.

The evidence in favour of replicability becomes weaker when adjusting bond portfolio returns by their exposure to market risk, proxied by the bond market portfolio (e.g., Dickerson et al., 2023b). For instance, when considering market-adjusted returns, only two portfolios show significant excess returns: a bond momentum factor constructed from equity momentum and a bond portfolio based on earnings announcement drift.

The results on bond risk factors expand on the evidence on equity anomalies by showing that by calculating bond returns in excess of their duration-matched US treasury bond rates produce a less optimistic view on risk factors replicability. This adds to the evidence on raw returns vs alpha hypothesis testing, which is opposite to equity portfolios.

5 Comparison with other methods to control FDR

The results on equity portfolio anomalies and bond risk factors suggest that replication rates can differ based on portfolio compositions, returns calculation, and sample size. Nonetheless, the overall picture of equity anomaly replicability is more optimistic than prior literature suggests (e.g., Hou et al., 2020). The evidence for corporate bond risk factors is less optimistic, with only a handful of risk factors which generate statistically significant risk-adjusted returns when controlling for FDR. These findings emphasize the difficulty of identifying true anomalies with limited time-series observations due to the noise in estimates of average raw and risk-adjusted returns.

Similar to the simulation study in Section 3, we now compare our `DDB` approach and its more aggressive variant (`DDBA`) to a range of baseline and widely used methods. These include linear step-up methods (`BH`, `BY`, `BKY`, `Storey`, `STS`), the bootstrap approach of Yekutieli and Benjamini (1999) (`YB`), and dependence-adjusted methods proposed by Fithian and Lei (2020) (`dBH`, `dBY`).

Table 2 reports the results for the Jensen et al. (2023) dataset and the mixed EW/VW

		Capped VW				VW			
		Raw returns		CAPM alpha		Raw returns		CAPM alpha	
		Thr(%)	Rep(%)	Thr(%)	Rep(%)	Thr(%)	Rep(%)	Thr(%)	Rep(%)
LSU	BH	1.96	41.63	2.92	68.63	1.23	26.42	2.69	56.28
	BKY	3.73	48.59	10.98	77.11	1.81	29.58	6.26	67.98
	BY	0.20	24.12	0.49	57.31	0.06	7.97	0.35	41.21
	Storey	6.23	53.59	19.96	83.06	3.69	35.69	13.41	73.54
	STS	91.65	100.00	92.36	100.00	3.93	36.37	25.33	79.56
Bootstrap	YB	0.78	32.82	1.36	63.33	0.37	15.36	1.10	47.63
	dBH	3.45	47.60	2.93	68.63	1.28	26.76	2.75	56.51
	dBY	8.33	55.76	2.65	67.75	0.38	15.56	1.74	51.04
Double Boot	DDB	2.34	47.7	9.49	75.16	1.75	28.79	6.49	67.97
	DDBA	1.19	35.78	20.50	82.67	4.46	37.25	15.32	75.29

		EW				Chen and Zimmermann (2021) sample			
		Raw returns		CAPM alpha		Raw returns		CAPM alpha	
		Thr(%)	Rep(%)	Thr(%)	Rep(%)	Thr(%)	Rep(%)	Thr(%)	Rep(%)
LSU	BH	3.27	67.52	2.92	80.39	3.31	69.48	3.63	77.03
	BKY	10.61	76.90	20.98	88.82	11.96	82.38	17.16	87.89
	BY	0.47	57.54	0.63	74.49	0.41	50.15	0.56	70.39
	Storey	18.64	82.42	38.24	93.61	38.83	93.03	34.83	92.11
	STS	23.54	84.39	89.00	100.00	84.82	100.00	77.09	98.82
Bootstrap	YB	1.46	64.57	1.68	77.80	1.26	58.33	1.55	72.88
	dBH	3.31	67.65	2.92	80.39	3.34	69.56	3.68	77.14
	dBY	2.39	65.39	2.92	80.39	2.49	64.97	3.08	76.18
Double Boot	DDB	8.53	75.16	16.18	86.93	12.87	82.41	17.59	87.05
	DDBA	19.13	81.92	34.07	92.42	28.68	89.49	36.87	92.56

Table 2: **MHT comparison for equity anomaly portfolios**. The table reports results for equal-weight (EW), value-weight (VW), and capped value-weight (Capped VW) portfolios from the Jensen et al. (2023) dataset, as well as the combined EW and VW portfolio sample from Chen and Zimmermann (2021). Results are shown for both raw returns and CAPM alphas, with replication rates (% Rep) and FDR thresholds (% Thr-p) presented for each method.

sample from Chen and Zimmermann (2021). Several key findings emerge. Across both raw returns and CAPM alphas, DDB achieves higher replication rates than conservative methods (e.g., BY, YB) while maintaining stricter FDR control than lenient methods (e.g., Storey, STS). For instance, for CAPM alphas, DDB yields replication rates of 75.6%, 67.9%, and 87.1% for Capped VW, VW, and EW portfolios, respectively, outperforming BH (68.6%, 56.3%, 80.4%) and dBH (68.6%, 56.5%, 80.4%).

Storey-based methods yield the highest replication rates (e.g., 100% for Capped VW and EW portfolios) but use lenient thresholds (e.g., p-values as high as 19.96% for Capped VW CAPM alpha), inflating the likelihood of false positives, as shown in Table 1. Conversely, `YB` and `dBY` apply stricter thresholds, leading to lower replication rates (e.g., 63.3% and 67.7% for Capped VW CAPM alphas). These results highlight `DDB`'s ability to balance power and FDR control effectively.

Replication rates are consistently higher for EW portfolios, as Hou et al. (2020) also noted. Capped VW portfolios offer a middle ground, addressing the trade-offs between EW's small-cap bias and VW's large-cap concentration (Jensen et al., 2023). Moreover, CAPM alphas consistently yield higher replication rates than raw returns, underscoring their utility in reducing noise from market risk exposure. For example, `DDB` achieves replication rates of 41.9% (Capped VW), 29.2% (VW), and 74.9% (EW) for raw returns, compared to 75.6%, 67.9%, and 87.1% for CAPM alphas.

		Bond returns				Duration-adjusted Bond returns			
		Raw returns		CAPM alpha		Raw returns		CAPM alpha	
		Thr(%)	Rep(%)	Thr(%)	Rep(%)	Thr(%)	Rep(%)	Thr(%)	Rep(%)
LSU	`BH`	1.26	40.00	0.86	26.67	0.00	13.33	0.00	13.33
	`BKY`	2.42	46.67	0.86	26.67	0.00	13.33	0.00	13.33
	`BY`	0.00	13.33	0.20	20.00	0.00	13.33	0.00	13.33
	`Storey`	10.17	60.00	0.86	26.67	10.24	60.00	0.00	13.33
	`STS`	92.06	100.00	5.35	40.00	70.49	100.00	2.54	20.00
Bootstrap	`YB`	0.50	23.00	0.25	20.47	0.00	13.33	0.00	13.33
	`dBH`	2.42	46.67	0.86	26.67	0.00	13.33	0.00	13.33
	`dBY`	1.26	40.00	0.86	26.67	0.00	13.33	0.00	13.33
Double Boot	`DDB`	2.27	45.80	0.84	26.40	0.96	22.93	0.00	13.33
	`DDBA`	5.17	53.00	2.41	32.60	5.24	47.67	0.00	13.33

Table 3: **MHT comparison for bond risk factors**. The table reports results for bond returns and bond duration-adjusted returns as in Dickerson et al. (2023a). Results are shown for both raw and market risk-adjusted returns (CAPM alphas), with replication rates (% Rep) and FDR thresholds (% Thr-p) presented for each method.

Table 3 reports the model comparison results for the bond returns and duration-adjusted returns. Similar to equity portfolio anomalies, `DDB` achieves higher replication rates than con-

servative methods, such as BY and YB. For instance, for CAPM alphas, DDB yields replication rates of 45.8% and 26.4% for raw returns and CAPM alphas, respectively, outperforming BY (13.3% and 20%) and YB (23% and 20.47%).

Storey-based methods yield the highest replication rates (e.g., 100% for raw returns and duration-adjusted returns) but imply lenient thresholds (e.g., p-values as high as 70.49% for raw duration-adjusted returns), inflating the likelihood of false positives. These results confirm DDB's ability to balance power and FDR control more effectively compared to popular methods.

Across datasets, DDB demonstrates robust performance. In the mixed EW/VW sample from Chen and Zimmermann (2021), it achieves a replication rate of 88%, outperforming BH (77%), BY (70.4%), and YB (72.9%). For bond portfolios, DDB and DDBA produce a relatively more optimistic picture without compromising FDR control. This highlights DDB's adaptability to diverse data structures and asset classes.

5.1 Time-series length vs cross-sectional coverage

In multiple hypothesis testing with FDR control, the relative dimensions of time series length (T) versus cross-section size (N) significantly impact test performance. When $N >> T$, size distortions increase as test statistics deviate from asymptotic distributions, and cross-sectional dependence undermines traditional FDR methods like Benjamini and Hochberg (1995). Conversely, when $T >> N$, tests exhibit greater power and more reliable FDR control due to improved estimation of time-series parameters and better asymptotic properties. Longer time series improve the precision of estimates and reduce the likelihood of Type II errors (failing to reject a false null hypothesis).

This dimensional relationship is particularly critical when testing for weak signals, which are prevalent in the context of portfolio anomalies (e.g., Harvey and Liu, 2020). To explore this issue, we examine two alternative sample periods starting in January 1942 and January 2004 (see vertical dashed lines in the left panel of Figure 1). The sample starting in 2004 features a larger cross-section, particularly for the CZ dataset, but sacrifices significant time

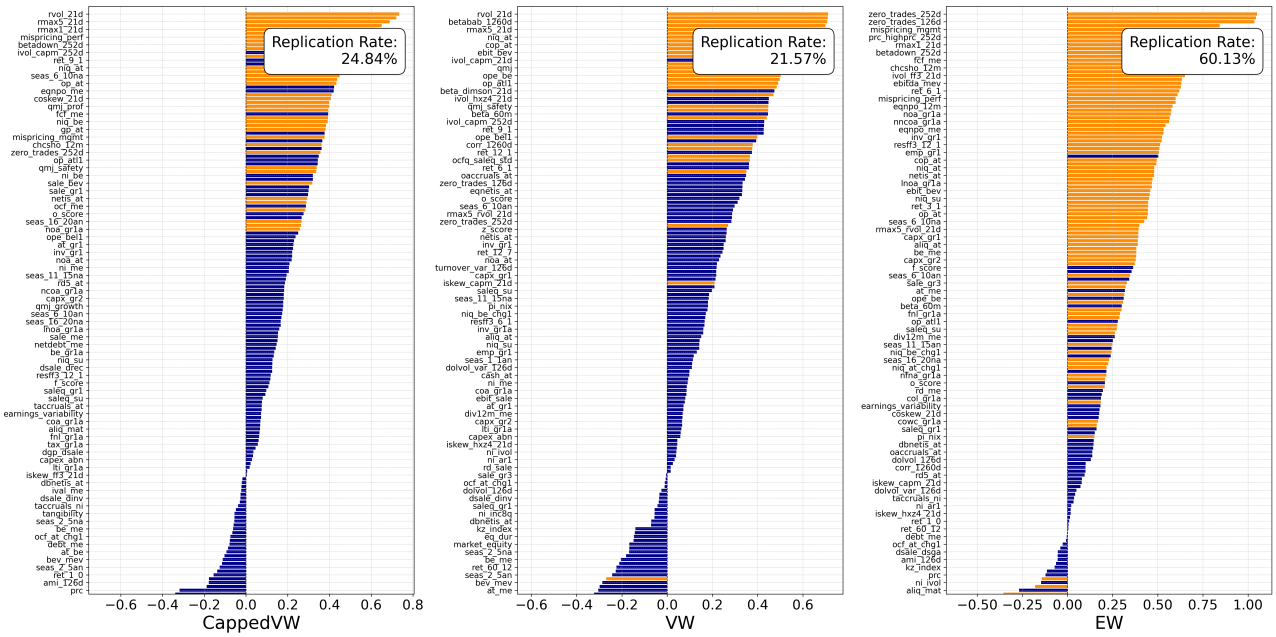


Figure 7: **Replication rates for shorter samples**. This figure reports the monthly CAPM alphas (in %) for different portfolio anomalies obtained from our DDB approach for a sample of portfolio anomalies continuously available since 2004. We report the results for capped value-weight, pure value-weight, and equal-weight portfolios. For ease of exposition, the y-axis in the subplots displays every other characteristic label.

series observations. Conversely, the sample starting in 1942 offers a longer time series but a smaller cross-section, which includes a long history of pre-publication performance for the selected anomalies, as shown in the right panel of Figure 1.

Figure 7 presents the replication rates (dark yellow), representing the percentage of portfolio anomalies with statistically significant CAPM alphas, for capped value-weight, pure value-weight, and equal-weight portfolios over the sample period starting in 2004. Not surprisingly, the replication rates drop significantly in the shorter sample. For example, the replication rate for capped VW portfolios decreases to 24.8% compared to 75.2% over the sample starting in January 1984. Similarly, the replication rate for equal-weight portfolios decreases from 86.9% to 60.13%, though it remains significantly higher than that for pure value-weight portfolios.

These results underscore the challenges of testing portfolio anomalies on shorter samples (e.g., Efron and Tibshirani, 1994). A reduced number of time-series observations amplifies noise in alpha estimates, increasing the variability of test statistics and making it harder to distinguish genuine signals from random variation. As a result, many anomalies fail to achieve

the significance threshold despite the larger cross-section of portfolio anomalies available in the post-2004 period.

Appendix D reports the replication results for a set of portfolio anomalies continuously available since January 1944. This sample represents a smaller cross-section compared to the main sample (see Figure 1). Figure D.5 confirms that longer time series provide more precise parameter estimates, increasing the likelihood of detecting true anomalies. Moreover, the smaller cross-section limits the total number of hypotheses tested, potentially reducing the burden of multiple testing.

For instance, the replication rate for CAPM alphas on capped value-weight portfolios increases to 89.4% (compared to 75.2% for the main sample). The gap with respect to EW portfolios (91.49%) becomes much smaller. Only a handful of characteristics related to trading frictions fail to achieve significance under FDR control. With more precise parameter estimates and fewer hypotheses, FDR control becomes more effective, leading to higher statistical power. However, the smaller cross-section may overlook newer anomalies, potentially biasing the results toward older, well-established signals (e.g., Green et al., 2017; Harvey, 2017).

6 Conclusion

This paper addresses a critical issue in empirical asset pricing: the challenge of identifying genuine portfolio anomalies amidst the vast cross-section of proposed characteristics and risk factors. By introducing a novel bootstrap method for multiple hypothesis testing (MHT), we provide a robust framework for controlling the false discovery rate (FDR) in the presence of complex correlation structures among portfolio anomalies. Through extensive simulations and empirical applications, we demonstrate that `DDB` achieves a superior balance between statistical power and FDR control, outperforming widely used methods such as the Benjamini-Hochberg (`BH`) and Benjamini-Yekutieli (`BY`) procedures. Importantly, our approach avoids the pitfalls of overly conservative or overly lenient corrections, making it a reliable tool for evaluating anomaly replicability.

Our empirical analysis reveals that the replicability of portfolio anomalies depends critically on portfolio construction strategies and the target of hypothesis testing. Equal-weight portfolios consistently achieve higher replication rates compared to value-weight portfolios, with capped value-weight schemes offering an intermediate solution. These findings highlight the influence of portfolio weighting on anomaly significance, as value-weight portfolios tend to concentrate on large-cap stocks, potentially obscuring signals from smaller firms. Moreover, CAPM-adjusted returns provide higher replication rates than raw returns, underscoring the importance of accounting for market risk when testing for anomalies. These results align with prior studies, such as Jensen et al. (2023), which emphasize the limitations of raw returns in identifying true anomalies relative to theoretical benchmarks.

Finally, we validate the robustness of our findings through additional tests on alternative datasets, including the characteristic-managed portfolios from Chen and Zimmermann (2021), and by examining shorter and longer samples with varying cross-sections. The results consistently show that `DDB` achieves reliable FDR control while maintaining high statistical power, even in settings with strong dependencies among hypotheses. This paper contributes to the growing literature on anomaly testing by offering a methodologically rigorous and empirically validated framework for assessing replicability.

References

Amihud, Y. (2002). Illiquidity and stock returns: cross-section and time-series effects. *Journal of financial markets*, 5(1):31–56.

Avramov, D., Cheng, S., and Metzker, L. (2023). Machine learning vs. economic restrictions: Evidence from stock return predictability. *Management Science*, 69(5):2587–2619.

Bai, J. and Ng, S. (2010). Instrumental variable estimation in a data rich environment. *Econometric Theory*, 26(6):1577–1606.

Barras, L., Scaillet, O., and Wermers, R. (2010). False discoveries in mutual fund performance: Measuring luck in estimated alphas. *The 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 (Methodological)*, 57(1):289–300.

Benjamini, Y., Krieger, A. M., and Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93(3):491–507.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.*, 29(4):1165–1188.

Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of finance*, 52(1):57–82.

Casella, G. and Berger, R. (2024). *Statistical inference*. CRC Press.

Chen, A. Y. and Dim, C. (2023). High-throughput asset pricing. *Working Paper*.

Chen, A. Y. and Zimmermann, T. (2021). Open source cross-sectional asset pricing. *Critical Finance Review, Forthcoming*.

Chen, A. Y. and Zimmermann, T. (2022). Publication bias in asset pricing research. *Working Paper*.

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

Dickerson, A., Julliard, C., and Mueller, P. (2023a). The co-pricing factor zoo. *working paper*.

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

Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. Chapman and Hall/CRC.

Fama, E. F. and French, K. R. (2010). Luck versus skill in the cross-section of mutual fund returns. *Journal of Finance*, 65(5):1915–1947.

Fithian, W. and Lei, L. (2020). Conditional calibration for false discovery rate control under dependence.

Genovese, C. (2004). A tutorial on false discovery control. https://www.stat.cmu.edu/~genovese/talks/hannover1-04.pdf. Accessed on 9/4/2021.

Green, J., Hand, J. R., and Zhang, X. F. (2013). The supraview of return predictive signals. *Review of Accounting Studies*, 18:692–730.

Green, J., Hand, J. R., and Zhang, X. F. (2017). The characteristics that provide independent information about average us monthly stock returns. *The Review of Financial Studies*, 30(12):4389–4436.

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

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

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

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

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

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

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

Kosowski, R., Timmermann, A., Wermers, R., and White, H. (2006). Can mutual fund "stars" really pick stocks? new evidence from a bootstrap analysis. *The Journal of finance*, 61(6):2551–2595.

Romano, J. P., Shaikh, A. M., and Wolf, M. (2008). Control of the False Discovery Rate under Dependence using the Bootstrap and Subsampling. *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research (invited article)*, 17(3):417–442.

Schwartzman, A. and Lin, X. (2011). The effect of correlation in false discovery rate estimation. *Biometrika*, 98(1):199–214.

Storey, J. D. (2002). A direct approach to false discovery rates. *JRSS: Series B*, 64(3):479–498.

Storey, J. D., Taylor, J. E., and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *JRSS: Series B*, 66(1):187–205.

van Binsbergen, J. H., Nozawa, Y., and Schwert, M. (2025). Duration-based valuation of corporate bonds. *The Review of Financial Studies*, 38(1):158–191.

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

Yekutieli, D. and Benjamini, Y. (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference*, 82(1):171 – 196.

Supplementary appendix for:

Correlated False Discoveries and Asset Pricing Anomalies

In this appendix, we provide a more detailed description of the data descriptive statistics for different portfolio configurations. We also provide complete derivations of the bootstrap approach proposed and its asymptotic properties. Finally, we also provide additional simulation and empirical results.

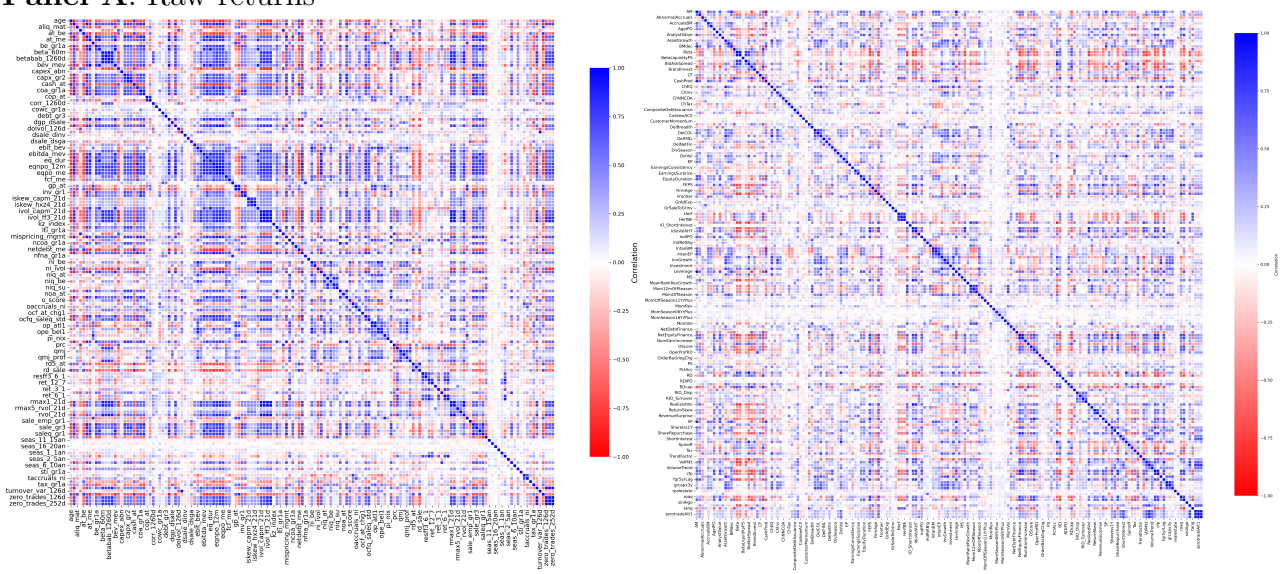
A Data descriptives

In this Section, we report some preliminary evidence on the cross-sectional correlation of the portfolio anomalies used in the main empirical analysis. Figure A.8 reports the heatmap of the cross-sectional correlation of the 153 portfolio anomalies from JKP (left panel) and the 176 portfolio anomalies from CZ (right panel) used in Section 4.1. For the sake of brevity, we report the capped value-weight construction for the JKP anomalies. As the panel of portfolios return is unbalanced for both cases (see Figure 1 in the main text), the pairwise correlations are calculated by considering only periods with overlapping returns. A color gradient from red (negative correlations) through white (zero correlation) to blue (positive correlations)

The correlation heatmap exhibits pronounced heterogeneity. For instance, for the JKP portfolios there is evidence of a group structure that align with fundamental investment dimensions. Return-based measures segregate into their own highly correlated cluster. A separate volatility and risk cluster displays strong positive intra-group correlations but predominantly negative relationships with return measures. Value metrics form another identifiable block with inverse relationships to the growth indicator cluster, highlighting the fundamental style dichotomy in asset pricing. Quality factors demonstrate exceptionally strong within-group correlations while maintaining distinctive cross-group relationships—typically positive with value indicators

but negative with volatility measures. Liquidity metrics constitute a tight correlation cluster largely isolated from other factors, suggesting they capture a unique market dimension. A similar heterogeneity arises when looking at the original portfolio constructions as reported by Chen and Zimmermann (2021) (right panel). For example, there are noticeable correlations within momentum factors.

Panel A: Raw returns



Panel B: Risk-adjusted returns (CAPM residuals)

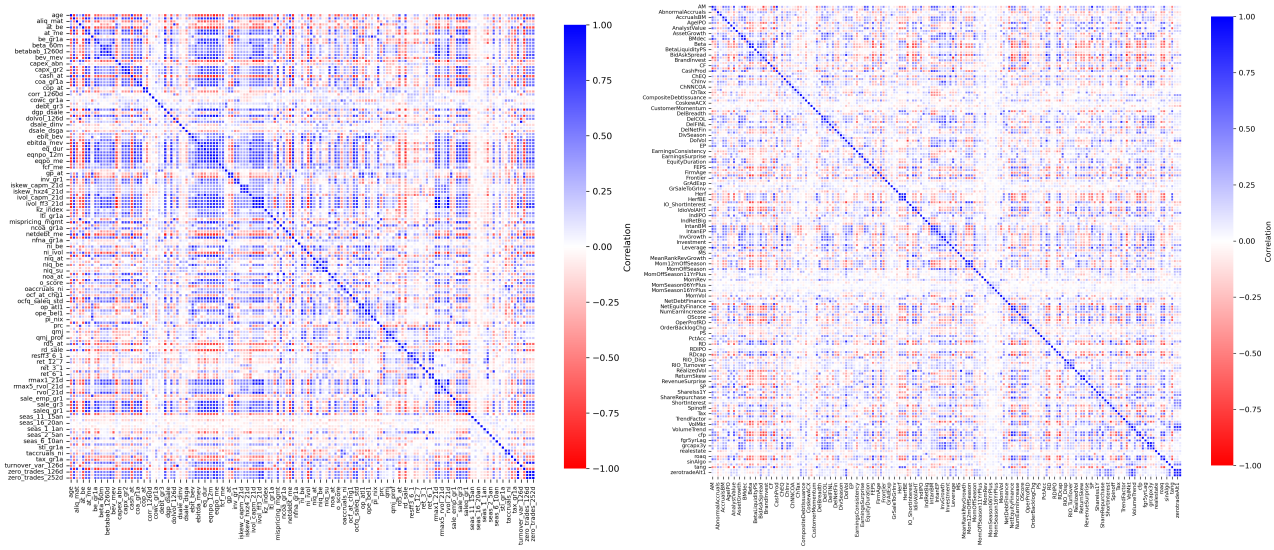
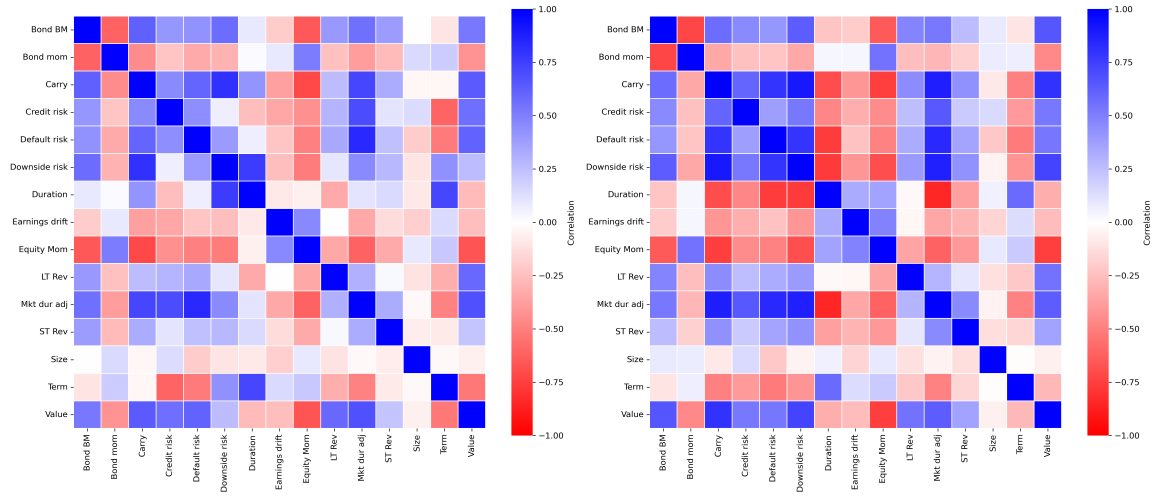


Figure A.8: **Equity portfolio returns correlation**. This figure reports the correlation heatmap for the 153 portfolio anomalies with capped value-weights from Jensen et al. (2023) (left panel) and the 176 portfolio anomalies as reported by Chen and Zimmermann (2021) (right panel). Panel A reports the correlation of the raw returns, whereas Panel B reports the correlation of the CAPM residuals, i.e., market-adjusted returns.

Panel A of Figure A.8 suggests that there are groups of economically relevant correlations when looking at raw portfolio returns. This reduces the effective number of independent tests, causing many rejections to reflect shared signals rather than distinct discoveries (e.g., Harvey et al., 2016). The correlation structure remains economically relevant even when accounting for the exposure to market risk, i.e., CAPM residuals, as illustrated in Panel B. This holds especially for the capped value-weight portfolios, whereas correlations are more modest, on average, for the Chen and Zimmermann (2021) portfolios.

Panel A: Raw returns



Panel B: Risk-adjusted returns (CAPM residuals)

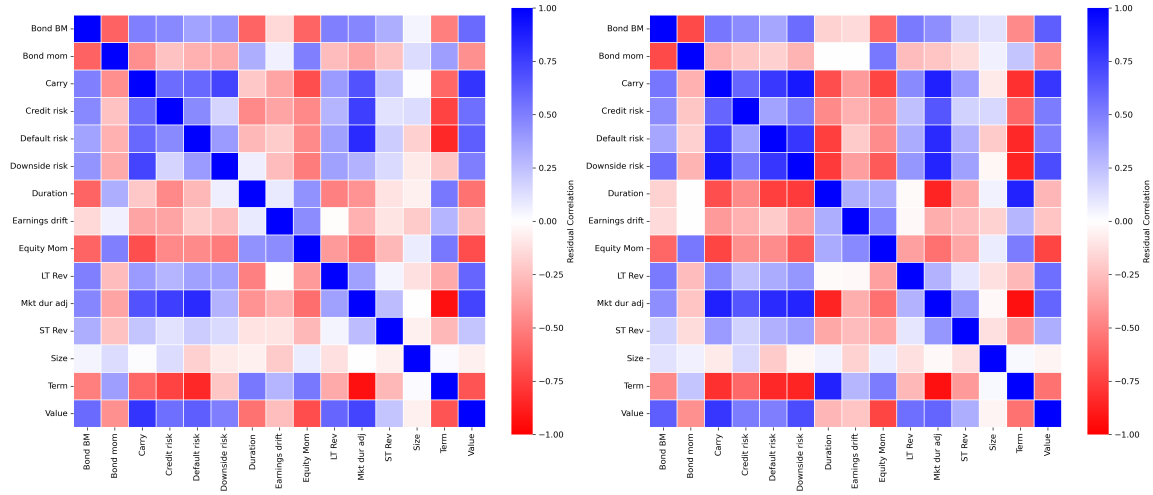


Figure A.9: **Bond returns correlation**. This figure reports the correlation heatmap for the 15 bond portfolio returns (left panel) and duration-adjusted returns (right panel) as used in the main empirical analysis. Panel A reports the correlation of the raw returns, whereas Panel B reports the correlation of the CAPM residuals, i.e., market-adjusted returns.

Figure A.9 reports the pairwise return correlations of the bond portfolio returns as from Dickerson et al. (2023a). Panel A reports the correlation for the raw strategy returns. The left panel shows the excess portfolio returns whereas the right panel reports the duration-adjusted excess returns (see van Binsbergen et al., 2025). Similar to equity portfolios, bond strategies show a quite heterogenous correlation structure, with carry, credit risk, default risk, and downrisk highly correlated to each other. Interestingly, the correlation increases for duration-adjusted returns. The duration-adjusted bond market portfolio (Mkt. dur adj) and the value factors are also highly correlated with carry, credit risk, default risk, and downside risk factors.

Panel B reports the correlation heatmap for bond excess returns (left panel) and duration-adjusted excess returns (right panel) when adjusting for the exposure to the bond market portfolio, i.e. CAPM residuals. Except for a few nuances, the heterogeneous correlation structure is largely consistent with raw returns. Thus, adjusting for a strong common factor, namely the market, does not mitigate the possibility of correlated hypotheses across bond factors.

B Proofs of theorems and lemmas

Proof of Theorem 1. For each simulation $b = 1, \ldots, B$, we sample (with replacement) T time steps $\mathcal{T}_b = (t_1^{(b)}, t_2^{(b)}, \ldots, t_T^{(b)}) \in \mathcal{S}(T)$, with $\mathcal{S}(T) = [T]^T$ be the set of all possible bootstrap samples of size T. Let

$$\widehat{e}_i^{(b)} = \frac{1}{T} \sum_{t \in \mathcal{T}_b} \widehat{\varepsilon}_{i,t}, \tag{B.1}$$

Using this, we calculate the bootstrap expectation of the resampled residuals. Since $\hat{e}_i^{(b)}$ is the average of resampled residuals, we have:

$$\mathbb{E}_{\text{boot}}[\widehat{e}_i^{(b)}] = \frac{1}{|\mathcal{S}(T)|} \sum_{\mathcal{T}_b \in \mathcal{S}(T)} \left(\frac{1}{T} \sum_{t \in \mathcal{T}_b} \widehat{\varepsilon}_{i,t} \right) \tag{B.2}$$

$$= \frac{1}{T} \sum_{t=1}^{T} \widehat{\varepsilon}_{i,t} \quad \text{(each } t \text{ appears equally in the bootstrap samples)} \tag{B.3}$$

$$= 0 \quad \text{(by least squares with intercept term).} \tag{B.4}$$

We next leverage the definition of covariance and the i.i.d. property of sampling with replacement:

$$T\mathrm{Cov}_{\mathrm{boot}}(\widehat{e}_i^{(b)}, \widehat{e}_j^{(b)}) = \frac{1}{T}\mathbb{E}_{\mathrm{boot}}\left[\left(\sum_{t\in\mathcal{T}_b}\widehat{\varepsilon}_{i,t}\right)\left(\sum_{t\in\mathcal{T}_b}\widehat{\varepsilon}_{j,t}\right)\right] \tag{B.5}$$

$$= \frac{1}{T}\mathbb{E}_{\mathrm{boot}}\left[\sum_{t\in\mathcal{T}_b}\widehat{\varepsilon}_{i,t}\widehat{\varepsilon}_{j,t}\right] \tag{B.6}$$

$$= \frac{1}{T}\sum_{t\in[T]}\widehat{\varepsilon}_{i,t}\widehat{\varepsilon}_{j,t}, \tag{B.7}$$

To approximate this covariance, we note that $\widehat{\varepsilon}_{i,t} - \varepsilon_{i,t} = \boldsymbol{x}_t^\top(\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i) \le |\boldsymbol{x}_t||\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i|$, where $\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i \sim \mathcal{N}(0, \sigma_i^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1})$. Using standard concentration bounds, for a fixed t,

$$|\widehat{\varepsilon}_{i,t} - \varepsilon_{i,t}| = O\left(\frac{\log(1/\delta)}{\sqrt{T}}\right) \quad \text{with probability at least } 1-\delta. \tag{B.8}$$

Since $\widehat{e}_i^{(b)}$ is the bootstrap average of T samples from $(\widehat{\varepsilon}_{i,t})_{t\in[T]}$, and $\widehat{\varepsilon}_{i,t}$ is zero-mean, the approximation becomes:

$$\frac{1}{T}\sum_{t\in[T]}\widehat{\varepsilon}_{i,t}\widehat{\varepsilon}_{j,t} = \mathrm{Cov}(\varepsilon_{i,1}, \varepsilon_{j,1}) + O\left(\frac{\log(1/\delta)}{\sqrt{T}}\right). \tag{B.9}$$

Thus,

$$T\mathrm{Cov}_{\mathrm{boot}}(\widehat{e}_i^{(b)}, \widehat{e}_j^{(b)}) = \mathrm{Cov}(\varepsilon_{i,1}, \varepsilon_{j,1}) + O\left(\frac{\log(1/\delta)}{\sqrt{T}}\right), \tag{B.10}$$

which concludes Theorem 1. □

Proof of Lemma 2. Define the following indicators for $\boldsymbol{u}', \boldsymbol{u}'' \in \mathbb{R}^N$:

$$m_{\mathrm{one}}(\boldsymbol{u}', \boldsymbol{u}'') = \sum_{i\in\mathcal{H}_0}\mathbf{1}\left[u_i' \ge u_i''\right], \tag{B.11}$$

$$m_{\text{two}}(\boldsymbol{u}', \boldsymbol{u}'') = \sum_{i \in \mathcal{H}_0} \mathbf{1}\left[|u_i'| \geq |u_i''|\right]. \tag{B.12}$$

It is straightforward to verify that for $m \in \{m_{\text{one}}, m_{\text{two}}\}$, we always have:

$$m(\boldsymbol{u}', \boldsymbol{u}'') \geq N_0/2 \quad \text{or} \quad m(\boldsymbol{u}'', \boldsymbol{u}') \geq N_0/2.$$

Define the event $\mathcal{D}(\boldsymbol{u}', \boldsymbol{u}'') = \{m(\boldsymbol{u}', \boldsymbol{u}'') \geq N_0/2\}$. This can be interpreted as a "duel" between two players $\boldsymbol{u}'$ and $\boldsymbol{u}''$, where at least one wins. To prove the lemma, we bound the probability of encountering a "hard-to-beat" random player $\boldsymbol{u}'$. Define the "skill" of $\boldsymbol{u}'$ as:

$$s(\boldsymbol{u}') = \Pr_{\boldsymbol{u}'' \sim \mathcal{U}}[\mathcal{D}(\boldsymbol{u}'', \boldsymbol{u}')]. \tag{B.13}$$

Let:

$$\mathcal{S}(q) = \{\boldsymbol{u}' \in \mathcal{U} : s(\boldsymbol{u}') \leq q\}, \tag{B.14}$$

$$S(q) = \Pr_{\boldsymbol{u}' \sim \mathcal{U}}[\boldsymbol{u}' \in \mathcal{S}(q)]. \tag{B.15}$$

We show that $S(q) \leq 2q$ for any $q \in (0, 1)$. Assume for contradiction that $S(q) = 2q + \varepsilon$ for some $q, \varepsilon > 0$. Then, by definition of $\mathcal{S}(q)$, for $\boldsymbol{u}' \in \mathcal{S}(q)$:

$$q \geq \int \mathbf{1}[\mathcal{D}(\boldsymbol{u}'', \boldsymbol{u}')] df(\boldsymbol{u}'') \tag{B.16}$$

$$\geq \int \mathbf{1}[\mathcal{D}(\boldsymbol{u}'', \boldsymbol{u}'), \boldsymbol{u}'' \in \mathcal{S}(q)] df(\boldsymbol{u}''). \tag{B.17}$$

Thus:

$$(2q + \varepsilon)q = q \int \mathbf{1}[\boldsymbol{u}' \in \mathcal{S}(q)] df(\boldsymbol{u}') \tag{B.18}$$

$$\geq \int \int \mathbf{1}[\mathcal{D}(\boldsymbol{u}'', \boldsymbol{u}'), \boldsymbol{u}', \boldsymbol{u}'' \in \mathcal{S}(q)] df(\boldsymbol{u}'') df(\boldsymbol{u}') \tag{B.19}$$

$$= \frac{(2q + \varepsilon)^2}{2}, \tag{B.20}$$

where the last equality holds because either $\mathcal{D}(\boldsymbol{u}'', \boldsymbol{u}')$ or $\mathcal{D}(\boldsymbol{u}', \boldsymbol{u}'')$ must always hold. This leads to a contradiction since $(2q + \varepsilon)q < \frac{(2q+\varepsilon)^2}{2}$ for $q, \varepsilon > 0$.

Now consider the complementary event $\mathcal{D}^c$. The probability of interest is:

$$1 - \Pr\left[\bigcup_{b\in[B]} \{m(\boldsymbol{u}^{(b)}, \hat{\boldsymbol{\alpha}}) \geq N_0/2\}\right] \tag{B.21}$$

$$\leq 1 - \Pr\left[\bigcup_{b\in[B]} \{m(\boldsymbol{u}^{(b)}, \boldsymbol{u}) \geq N_0/2\}\right] \tag{B.22}$$

$$= \Pr\left[\bigcap_{b\in[B]} \{m(\boldsymbol{u}^{(b)}, \boldsymbol{u}) < N_0/2\}\right] \tag{B.23}$$

$$= \Pr\left[\bigcap_{b\in[B]} \{\mathcal{D}^c(\boldsymbol{u}^{(b)}, \boldsymbol{u})\}\right]. \tag{B.24}$$

By independence of bootstrap samples:

$$\Pr\left[\bigcap_{b\in[B]} \{\mathcal{D}^c(\boldsymbol{u}^{(b)}, \boldsymbol{u})\}\right] \leq \int_0^1 (1-q)^B dS(q) \tag{B.25}$$

$$\leq \int_0^{1/2} (1-q)^B 2q dq \tag{B.26}$$

$$= \frac{2}{B+1}[(1-q)^{B+1}]_0^{1/2} \tag{B.27}$$

$$\leq \frac{2}{B+1}. \tag{B.28}$$

This establishes the desired result, completing the proof. □

The following Lemma provides the error of the empirical FDR.

Lemma 4. (Accuracy of the empirical FDR) *For each* $\hat{\boldsymbol{\alpha}}^{(b)}$ *and* $\delta \in (0, 1)$*, with probability at least* $1 - 2\delta$*, we have*

$$|\mathrm{FDR}(\widehat{\boldsymbol{\alpha}}^{(b)}, c_q) - \widehat{FDR}(\widehat{\boldsymbol{\alpha}}^{(b)}, c_q)| \leq \sqrt{\frac{\log(1/\delta)}{2B}}. \tag{B.29}$$

The proof of Lemma 4 follows directly from Hoeffding's inequality. Since $(\widehat{\mathrm{FDR}}^{(b)})_{b=1,2,\ldots,B}$

are unbiased estimators of FDR, bounded in $[0,1]$, and independent of each other, the bound holds. Finally, the optimal threshold c_q can be obtained via binary search such that:

$$\inf_{c_q \in (0,1)} \widehat{\mathrm{FDR}}(\hat{\boldsymbol{\alpha}}^{(b)}, c_q) \leq q/2,$$

assuming the monotonicity of FDR. We now use Lemma 4 to prove the theoretical convergence of the FDR estimate.

Proof of Theorem 3. Let $e = \sqrt{\frac{\log((BS)/\delta)}{2W}}$. By applying a union bound to Lemma 4 with $\delta = \delta/(BS)$ over all $b = 1, 2, \ldots, B$ and S candidates $c_q^{(b,s)}$ of $c_q^{(b)}$, we have:

$$\forall b \in [B], s \in [S]: \ |\mathrm{FDR}(\widehat{\boldsymbol{\alpha}}^{(b)}, c_q^{(b,s)}) - \widehat{\mathrm{FDR}}(\widehat{\boldsymbol{\alpha}}^{(b)}, c_q^{(b,s)})| \leq e, \tag{B.30}$$

with probability at least $1 - 2\delta$. Define:

$$C_q(e) = \min_{b \in [B]} \sup\{c_q \in (0,1) : \mathrm{FDR}(\widehat{\boldsymbol{\alpha}}^{(b)}, c_q) \leq q/2 + e\}, \tag{B.31}$$

$$c_q^*(e) = \sup\{c_q \in (0,1) : \mathrm{FDR}(\boldsymbol{\alpha}, c_q) \leq q + 2e\}. \tag{B.32}$$

Using Eq.(B.30), we derive:

$$\min_{b \in [B]} c_q^{(b)} \leq C_q(e) \tag{B.33}$$

$$\leq c_q^*(e), \tag{B.34}$$

where the second inequality uses Eq.(11) with $q' = q + 2e$. By the non-decreasing property of FDR, we have:

$$\mathrm{FDR}\left(\boldsymbol{\alpha}, \min_{b \in [B]} c_q^{(b)}\right) \leq \mathrm{FDR}(\boldsymbol{\alpha}, C_q(e)) \tag{B.35}$$

$$\leq \mathrm{FDR}(\boldsymbol{\alpha}, c_q^*(e)) = q + 2e. \tag{B.36}$$

This completes the proof. □

C Additional simulation results

C.1 FDR and p-value thresholds for shorter time series

Table C.1 reports the results for shorter time series ($T = 50$). The findings largely confirm the trends observed in Table 1. Our DDB and DDBA methods strike a robust balance between FDR control and power. While DDB is the safer choice for strict FDR control, DDBA provides an effective compromise between control and rejection rates, outperforming both restrictive and lenient methods across most scenarios.

DDB consistently achieves competitive FDR control across all settings. While its FDR is slightly higher than BH and BKY, it remains within acceptable limits, striking a balance between error control and power. DDB exhibits higher rejection rates than conservative methods (e.g., BH, BKY, BY), particularly at higher correlations ($\rho = 0.5, 0.9$). However, it is more conservative compared to liberal methods like STS and DDBA.

For $\pi_0 = 0.25$ DDB matches BH and BKY in threshold (0.07%) and rejection rates (2.2%) while maintaining slightly higher FDR (1.32%) at low correlation ($\rho = 0$). At high correlation ($\rho = 0.9$), DDB achieves superior FDR control (1.38%) compared to Storey (4.66%) and STS (8.42%) with moderate rejection rates (8.92%). For $\pi_0.5$, under low correlation ($\rho = 0$), DDB (FDR = 2.11%, Rej = 1.14%) competes well with BH and BKY, offering slightly higher rejection rates. At high correlation ($\rho = 0.9$), DDB maintains lower FDR (2.62%) than Storey (8.53%) and STS (15.33%), though with fewer rejections.

Overall, DDB balances FDR control and power effectively, making it reliable across diverse settings. DDB is particularly robust to high correlations and varying true null proportions. Specifically, DDB outperforms liberal methods like Storey and STS in FDR control and is more reliable than DDBA when strict error control is prioritized.

Panel A: True null ratio $\pi_0 = 0.25$

	$\rho = 0$			$\rho = 0.5$			$\rho = 0.9$		
Method	Thr(%)	FDR(%)	Rej(%)	Thr(%)	FDR(%)	Rej(%)	Thr(%)	FDR(%)	Rej(%)
Single	4.10	7.09	16.04	3.89	11.21	16.34	2.85	6.37	16.34
BH	0.07	1.20	2.20	0.20	1.10	4.76	0.28	0.82	6.52
BKY	0.07	1.19	2.20	0.43	1.28	5.60	0.73	1.37	8.12
BY	0.00	0.37	0.48	0.01	0.18	1.24	0.02	0.14	2.20
Storey	0.18	1.89	3.44	3.46	2.91	11.64	7.30	4.66	20.72
STS	15.61	5.98	20.64	20.03	6.93	27.42	21.45	8.42	35.00
YB	0.02	0.90	1.10	0.06	0.77	2.82	0.24	1.16	6.24
DDB	0.07	1.32	2.20	0.47	1.28	5.76	0.86	1.38	8.92
DDBA	0.37	2.40	4.90	1.35	2.47	9.50	1.87	2.36	13.00

Panel B: True null ratio $\pi_0 = 0.5$

	$\rho = 0$			$\rho = 0.5$			$\rho = 0.9$		
Method	Thr(%)	FDR(%)	Rej(%)	Thr(%)	FDR(%)	Rej(%)	Thr(%)	FDR(%)	Rej(%)
Single	3.87	19.15	12.26	3.64	21.87	12.56	2.70	9.39	12.56
BH	0.04	2.43	1.30	0.11	2.00	2.78	0.20	1.70	4.58
BKY	0.04	2.33	1.26	0.20	2.17	3.14	0.44	2.10	5.44
BY	0.00	0.62	0.34	0.01	0.53	0.66	0.01	0.38	1.40
Storey	0.07	3.13	1.70	1.88	4.78	6.96	5.77	8.53	16.54
STS	12.03	10.22	15.02	15.88	11.48	20.78	18.61	15.33	30.06
YB	0.01	1.73	0.72	0.04	1.33	1.62	0.17	1.80	4.36
DDB	0.03	2.11	1.14	0.24	2.16	3.22	0.54	2.62	6.20
DDBA	0.16	4.61	2.54	0.70	4.16	5.70	1.24	4.33	9.28

Panel C: True null $\pi_0 = 1$

	$\rho = 0$			$\rho = 0.5$			$\rho = 0.9$		
Method	Thr(%)	FDR(%)	Rej(%)	Thr(%)	FDR(%)	Rej(%)	Thr(%)	FDR(%)	Rej(%)
Single	3.13	91.70	4.84	2.28	61.30	5.00	0.86	19.25	4.76
BH	0.00	4.20	0.10	0.01	3.80	0.34	0.06	2.00	1.44
BKY	0.00	4.15	0.08	0.02	3.75	0.40	0.14	1.65	1.62
BY	0.00	0.75	0.02	0.00	0.80	0.02	0.00	0.25	0.16
Storey	0.00	4.65	0.10	1.21	8.20	3.30	5.98	17.25	17.16
STS	6.80	14.05	7.46	11.37	17.65	13.86	7.60	19.05	19.06
YB	0.00	2.60	0.06	0.00	3.35	0.12	0.03	2.90	0.86
DDB	0.00	3.05	0.06	0.06	4.25	0.70	0.38	3.80	3.44
DDBA	0.01	7.50	0.18	0.24	7.55	1.68	0.76	6.10	5.54

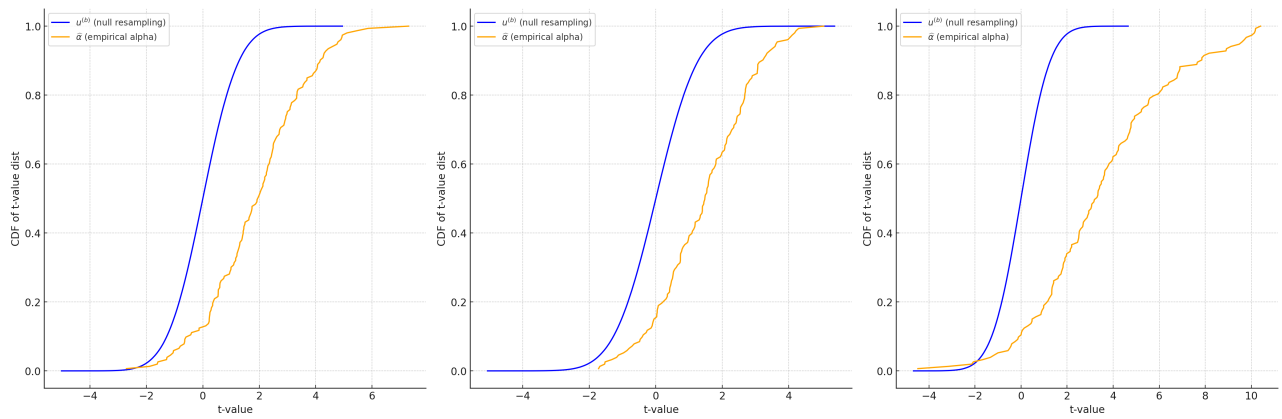
Table C.1: **Simulation results for shorter time series**. The table reports the simulation results for $T = 50$. The performance of each method is reported in terms of p-value threshold (Thr-p), proportion or rejected hypotheses (Rej) and estimated FDR. These are average values out of 2,000 simulation runs. The table reports the results for different level of hypotheses correlation ($\rho = 0, 0.5, 0.9$).

D Additional results

In this section, we present the cumulative distribution functions (CDFs) of the test statistics under the null hypothesis, generated through a double-bootstrap approach. These null dis-

tributions are centered around zero, reflecting the assumption that the null hypothesis holds in simulation. The null distributions are compared to the empirical test statistics derived from observed data, which may exhibit significant deviations from the null, depending on the presence of true effects. The degree of separation between the empirical alphas and the null distribution gives insights into the statistical significance of the test statistics.

Figure D.1 compares the bootstrap null CDF and the empirical t-statistic CDF for raw returns across various portfolio constructions (capped value-weight, pure value-weight, and equal-weight). The empirical alpha CDF (orange) is consistently shifted to the right relative to the null distribution CDF (blue), indicating that the empirical t-values are systematically larger than those expected under the null hypothesis. This suggests statistical evidence against the null hypothesis across all portfolio configurations.

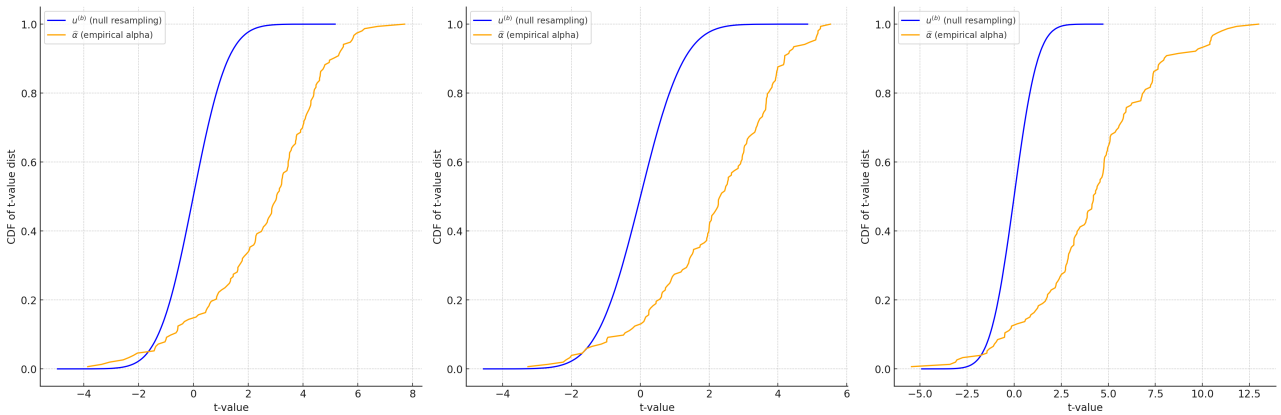


(a) CappedVW (b) VW (c) EW

Figure D.1: **Bootstrap vs true t-test distribution for raw returns**. Distribution of resampled residuals $(u_i^{(b)})_{i\in[N],b\in[B]}$ ("$u^{(b)}$ (null resampling)") and random values drawn from the true t-distribution ("empirical alpha") for the raw returns.

Figure D.2 examines the test statistics for CAPM alphas. Compared to the results for raw returns, the separation between the empirical alpha CDF (orange) and the null distribution CDF (blue) is noticeably broader. This broader range of empirical t-values for CAPM alphas reflects stronger deviations from the null hypothesis, suggesting that portfolio anomalies persist even after adjusting for risk factors.

These results underscore the statistical significance of the observed test statistics. If the null



(a) CappedVW (b) VW (c) EW

Figure D.2: **Bootstrap vs true t-test distribution for CAPM alphas**. Distribution of resampled residuals $(u_i^{(b)})_{i\in[N],b\in[B]}$ ("$u^{(b)}$ (null resampling)") and random values drawn from the true t-distribution ("empirical alpha") for the raw returns.

hypothesis were true, the empirical alpha t-values would closely align with the bootstrap null distribution. However, the observed shift indicates that the data are unlikely to conform to the null hypothesis, particularly for CAPM alphas. This highlights the robustness of portfolio anomalies when accounting for risk-adjusted returns.

D.1 Economic groups and replicability

Figure D.6 reports replication rates for portfolio anomalies grouped by their economic meaning as described in Chen and Zimmermann (2021).

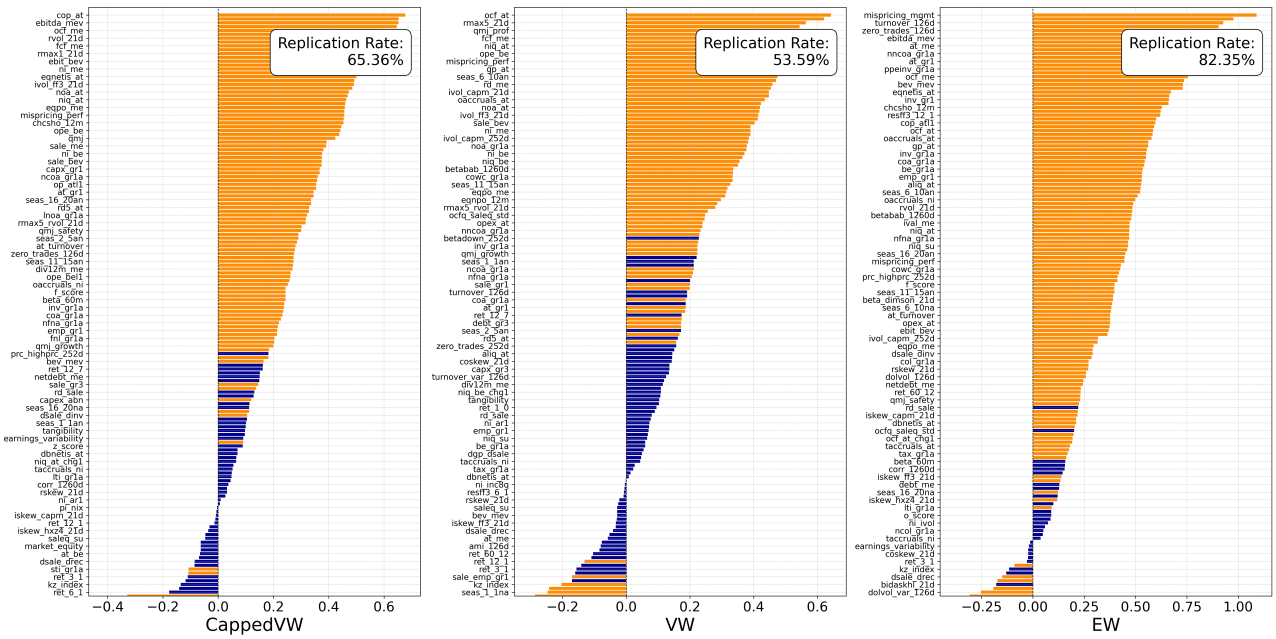


Figure D.3: **Replication rates for four-factor alphas**. This figure reports the monthly alphas (in %) from a four-factor model that include size, value, and momentum in addition to the market excess returns (Carhart, 1997). The anomalies that are successfully replicated according to our DDB approach are highlighted in dark yellow. The sample size is from January 1984 to December 2023. The left panel shows the results for the capped value-weight portfolios, the middle panel for the value-weight portfolios, and the right panel for the equal-weight portfolios.

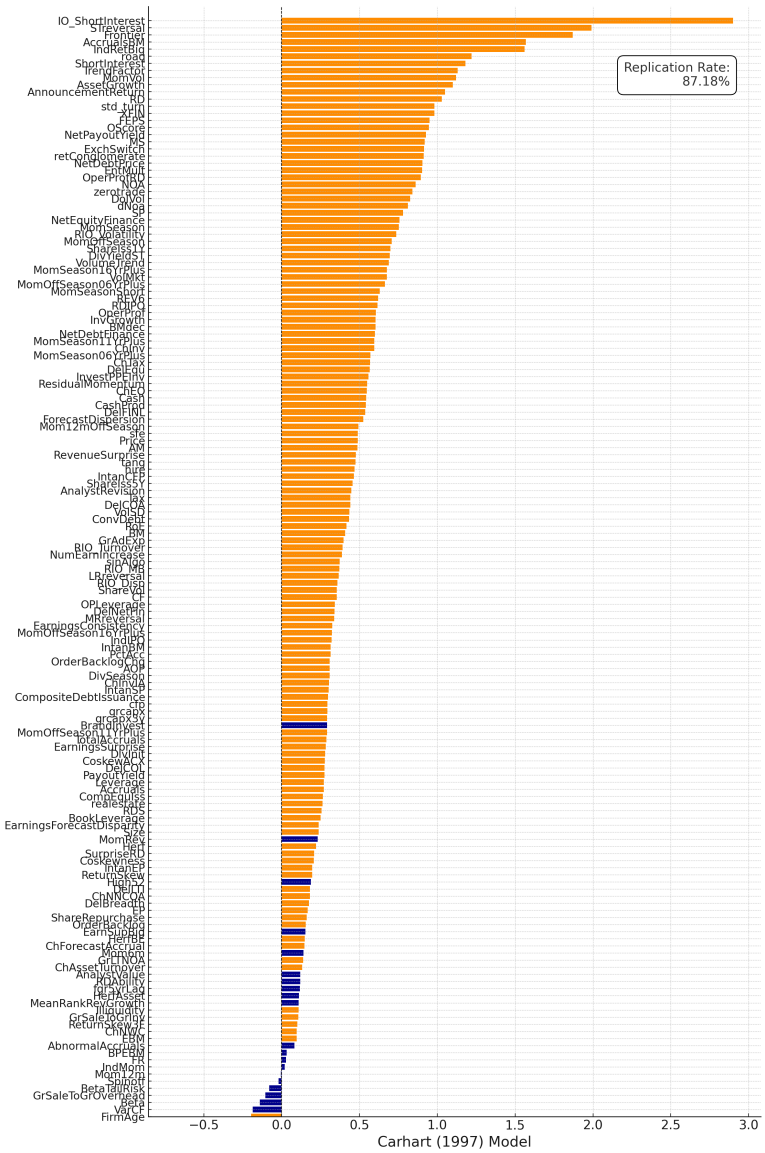


Figure D.4: **Replication rates for four-factor alphas**. This figure reports the monthly alphas (in %) from a four-factor model that include size, value, and momentum in addition to the market excess returns (Carhart, 1997). The anomalies that are successfully replicated according to our DDB approach are highlighted in dark yellow. The sample size is from January 1984 to December 2023. The left panel shows the results for the capped value-weight portfolios, the middle panel for the value-weight portfolios, and the right panel for the equal-weight portfolios.

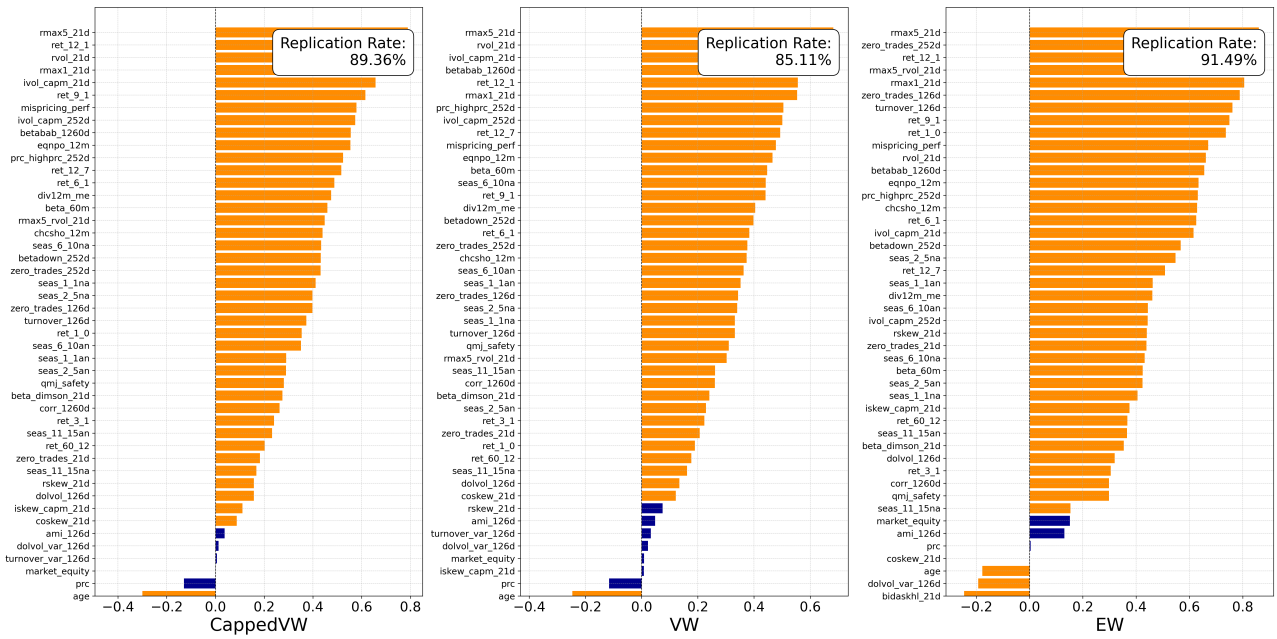


Figure D.5: **Replication rates for longer samples**. This figure reports the monthly CAPM alphas (in %) for different portfolio anomalies obtained from our DDB approach for a sample of portfolio anomalies continuously available since 1944. We report the results for capped value-weight, pure value-weight, and equal-weight portfolios.

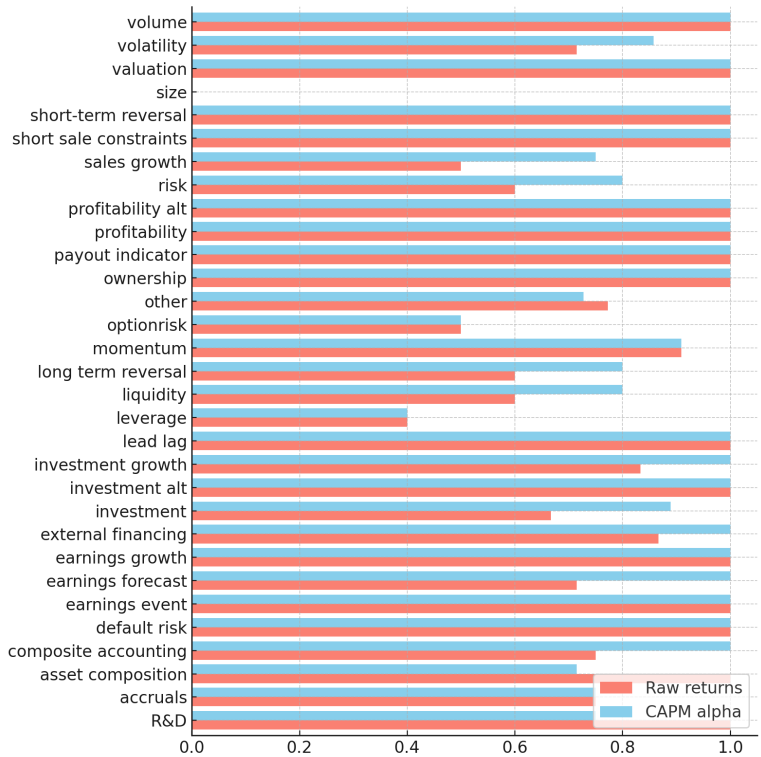


Figure D.6: **Replication rates by economic category**. This figure reports the replication rate (in decimals) for each anomaly clustered by economic category as in Chen and Zimmermann (2021). The replication rates for CAPM alphas are shown in light blue, and those for raw returns are shown in light red. The sample spans January 1984 to December 2022.