Factors or Fake? A New Look at Anomalies*

Siddhartha Chib Shuhua Xiao Lingxiao Zhao

March 2026

Abstract

Based on the insight that declaring an anomaly as not spanned requires knowledge of the true stochastic discount factor (SDF), we reverse conventional testing and suppose that, under the alternative, the anomaly is spanned. Under this novel framework, we formulate a procedure that computes posterior probabilities for single anomalies, pairs, and larger sets, integrating over the distribution of possible SDFs, while controlling the expected false discovery rate at each stage. In the U.S. data, 132 of 153 anomalies are classified as spanned. The evidence highlights redundancy and masking in the anomaly zoo.

Key Words: Bayesian multiple testing; Expected false discovery rate control; Model uncertainty; Posterior spanning probabilities; Spanned anomalies

JEL Classification: C11, C12, G10, G12

*We thank the seminar and conference participants at the Sun Yat-sen University, 4th Hong Kong Conference for Fintech, AI, and Big Data in Business, for their valuable comments. Chib (E-mail: chib@wustl.edu) is at Olin School of Business, Washington University in St. Louis; Xiao (E-mail: shxiao3-c@my.cityu.edu.hk) is at City University of Hong Kong; Zhao (E-mail: lingxiao@phbs.pku.edu.cn) is at Peking University HSBC Business School. Zhao acknowledges the support from the National Natural Science Foundation of China (Grant No. 72503015) and the Shenzhen Outstanding Talent Program for Scientific and Technological Innovation (Grant No. RCBS20231211090520033).

1 Introduction

Over the past two decades, the empirical asset pricing literature has seen a proliferation of reported return anomalies, cross-sectional patterns in average returns that appear unexplained by standard benchmark models (Cochrane, 2011; Harvey et al., 2016). This explosion of findings has sparked concerns about data mining, multiple testing, and replication (Harvey et al., 2016; McLean and Pontiff, 2016; Harvey and Liu, 2020), raising fundamental questions. Which anomalies represent distinct sources of priced risk? Which are redundant signals that can be explained by other factors?

In the conventional testing approach applied in finance to answer these questions, one starts from a benchmark model such as the CAPM (Sharpe, 1964; Lintner, 1965), the Fama–French three-, four-, or five-factor models (Fama and French, 1993; Carhart, 1997; Fama and French, 2015), or the q-factor model (Hou et al., 2015), and tests the null hypothesis that the given anomaly is spanned by the benchmark model. If one lets α denote the intercept in this regression, the test reduces to $H_0 \colon \alpha = 0$ against the alternative that α is different from zero, that is, the hypothesis that the anomaly is not spanned. It is common in finance to implement such a test by a t-test or by constructing a confidence interval for α. One practical difficulty with this approach is that it cannot be used to establish evidence in favor of the null hypothesis. Even if the t-test fails to reject the null, or the confidence interval fails to exclude zero, it is not possible within this framework to accept the null and conclude that the anomaly is spanned. Failure to reject does not mean that α is zero; the same confidence interval shows that many values close to zero are also consistent with the data, so there is no mechanism to verify that the intercept is truly zero.

A separate difficulty is that the test is conditional on the specified benchmark model. When the null is rejected, the conclusion is strictly limited to the finding that the anomaly is unspanned relative to that benchmark. One cannot use this rejection to infer that the anomaly represents a genuinely new dimension of risk, since it is quite

possible that the evidence against the null is an artifact of the assumed benchmark and would be reversed under a different one (Lewellen et al., 2010; Harvey, 2017). This benchmark dependence can have a non-trivial effect on the conclusion. One cannot attempt to address this by searching over possible factor combinations, since the search process (even the order in which the benchmarks are changed) alters the level of the test in an unknown way.

The practical severity of both difficulties is illustrated by our data. Among the 153 candidate anomalies in Jensen et al. (2023), 97 fail to meet the 5% significance threshold against `FF6`. Of these, 74 have confidence intervals wide enough to include a monthly alpha of 20 basis points, roughly 2.4% annualized. The insignificance of these anomalies is thus driven largely by estimation noise rather than by point estimates close to zero, and the test provides no basis for concluding that their intercepts are truly zero. The benchmark dependence problem is equally stark. Figure 1 reports the significance status of some anomalies across four standard benchmarks: CAPM, FF3, FF5, and FF6. The complete figure is reported in the appendix.

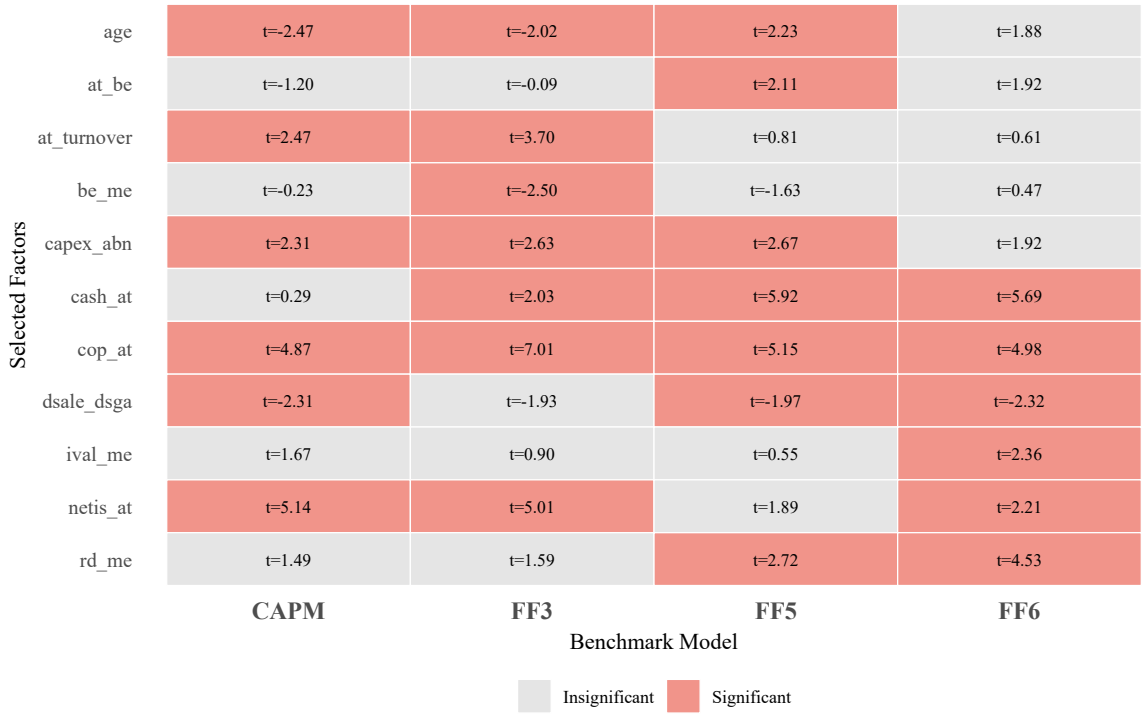
As the figure shows, the significance of anomalies flips across specifications, with factors that appear as strong alphas under one benchmark disappearing under another. Together, these patterns confirm that the two difficulties identified above are not merely theoretical concerns but are quantitatively important features of the data.

A related proposal is to raise the significance threshold, requiring t-statistics of 2.5 or 3 rather than the conventional 1.96 (Harvey et al., 2016). While this reduces the rate of false discoveries, it does not resolve either difficulty. A higher threshold makes it even harder to establish spanning from a failure to reject, and the conclusion upon rejection remains conditional on the chosen benchmark. Raising the bar for significance is a useful diagnostic, but it does not provide a mechanism for verifying redundancy or for making inferences that are robust to the choice of benchmark model.

In this paper, we recognize that these two difficulties make it nearly impossible to definitively establish whether an anomaly is genuinely new or merely redundant,

Figure 1 Sensitivity of Anomaly Significance to Benchmark Model Selection

Note: Statistical significance of some anomalies across four standard benchmark models: CAPM, FF3, FF5, and FF6. Significance is assessed by a t-test of $H_0 : \alpha = 0$ at the 5% level, where α is the intercept from a time-series regression of the anomaly return on the benchmark factors. Red tiles indicate rejection of the null at the 5% level; gray tiles indicate failure to reject.



and we introduce a new approach to anomaly evaluation that does not share these drawbacks. The approach distinguishes between spanned and unspanned anomalies by reversing the traditional hypothesis structure. Rather than placing the null on spanning, we take the null to be that the anomaly cannot be spanned by a given benchmark or any of its subsets, and the alternative (the discovery) corresponds to the hypothesis that it can be spanned by the given benchmark or one of its subsets, identifying it as a *fake anomaly*.

Before we explain why this inversion is the correct way to proceed, we emphasize that this inversion of hypotheses is new. It has not been discussed in the literature because, from the classical testing framework, the sampling distribution under the null becomes non-standard. But interestingly, these hypotheses can be tested within the

Bayesian framework. We have to modify the conventional logic of Bayesian testing to accommodate the specifics of the current problem, but the approach we develop is readily implementable and has theoretical guarantees.

Our motivation for proposing this new paradigm is simple. If a set of factors is found to span an anomaly, the redundancy of that anomaly is verified. This is because a factor that can be replicated by some combination of traded factors carries no independent pricing information, regardless of whether that combination is the best possible one. We also establish formally that if an anomaly is spanned by one benchmark, it remains spanned under any richer benchmark that contains it. Thus, a finding of spanning is conclusive and does not depend on having identified the optimal factor set. By shifting the burden of proof in this way, our inferences about spanning do not require access to the optimal benchmark.

The Bayesian approach we develop to test the inverted hypotheses is designed to handle a large number of anomalies. For any given anomaly, the alternative hypothesis corresponds to the claim that it can be spanned by some benchmark. To assess this, we use all admissible partitions of the anomaly and a starting benchmark, in the manner of Chib and Zeng (2020) and Chib et al. (2020), that split those components into *spanned* and *unspanned* parts. We then compute the total posterior probability of spanning by averaging over all splits in which the anomaly is spanned. We repeat this for each anomaly and apply an expected false discovery rate (EFDR) procedure to control for multiple testing. After eliminating anomalies that are spanned under this EFDR criterion, we repeat the process for anomalies considered in pairs, triplets, and larger sets. This sequential expansion is designed to uncover anomalies that appear unspanned in isolation but are revealed as redundant when considered jointly. We prove that our EFDR step-up rule at each stage preserves global EFDR control across the cumulative discovery set. Finally, we show that the spanning conclusions drawn within the observable payoff span are immune to omitted SDF components. Any component of the true SDF orthogonal to the working payoff space has no effect on the pricing or

spanning restrictions evaluated by the method, so an SDF outside the model universe cannot overturn the classifications.

We evaluate the finite-sample performance of our proposed method in an extensive Monte Carlo study in which the spanning status of each anomaly is known by construction. We use three designs to stress-test the procedure. A baseline with equal numbers of spanned and unspanned anomalies, a masking design in which two truly spanned anomalies are individually hard to classify but jointly recoverable, and a benchmark uncertainty design in which some anomalies are spanned only by a proper subset of `FF6`. Across all three designs and two signal regimes, Bayes EFDR substantially outperforms fixed-benchmark frequentist procedures and symmetric Bayes testing on the key recovery margins, and is the only procedure that reliably resolves masking.

In the empirical analysis, we apply the procedure to the 153-factor database of Jensen, Kelly, and Pedersen (2023) over a January 1985 through December 2024 sample, using `FF6` as the benchmark. The sequential Bayes–EFDR scan at $q = 0.20$ classifies 126 of the 153 candidate anomalies as fake, leaving 27 survivors. The identification path is highly nonuniform across steps: Steps 1 and 2 alone account for 108 fake anomalies, with Step 2 contributing 87 in a single pass. This concentration at Step 2 is direct evidence of masking. Entire economic themes, Investment, Momentum, and Low risk, are completely absorbed by `FF6`, while Size and Value are predominantly spanned, with 80% and 83% of their members classified as fake, respectively. The efficiency ratio of the Step-2 fake anomalies is only 0.04, confirming that the 87 factors discarded at that stage collectively span fewer than four independent dimensions; the procedure eliminates redundancy rather than pricing information.

We further document the masking effect through a counterfactual on model uncertainty. When model averaging is suppressed and spanning is evaluated against a restricted two-model space, the procedure classifies 58 anomalies as fake in Step 1 alone, compared with 21 under the full procedure. This difference of 37 anomalies represents

factors that can be linearly priced by a small benchmark but carry independent pricing information once the full distribution of SDFs is integrated out. The pattern repeats at every subsequent step: suppressing model uncertainty at any stage causes a discrete spike in spurious spanning classifications, confirming that the lower fake-anomaly counts in the baseline reflect a rigorous evidential standard rather than low statistical power.

To validate that the 27 survivors are economically distinct from the 126 fake anomalies, we conduct a series of out-of-sample comparisons across one million matched pairs of five-factor sets drawn from each group. Survivor sets consistently outperform fake-anomaly sets on log average predictive likelihood, out-of-sample Sharpe ratio, and cross-sectional pricing metrics across three test-asset universes. The combination of five survivors augmented with `FF6` achieves a mean out-of-sample Sharpe ratio of 1.18, compared with 0.69 for the matched fake-anomaly specification. Adding survivors to `FF6` also reduces RMS alpha and improves total R^2 relative to equally sized fake-anomaly augmentations across the Fama–French 49-industry, 100 P-Tree, and 120 bisort test portfolios.

To assess robustness to the initial benchmark choice, we rerun the procedure by replacing the five non-market `FF6` factors with the Hou, Xue, and Zhang (2020) profitability and investment factors, the Pástor and Stambaugh (2003) liquidity factor, the Frazzini and Pedersen (2014) betting-against-beta factor, and the Asness and Frazzini (2013) value factor, retaining only `Mkt` from the original benchmark. Despite the two benchmarks sharing only the market factor, the intersection of their fake-anomaly sets contains 116 elements, yielding a Jaccard index of 0.88 and an F_1 score of 0.94. Taking the union of the two fake-anomaly sets leaves 21 anomalies that survive under both benchmarks. These 21 factors organize into four economically interpretable channels that the benchmarks leave unspanned: fundamental quality (cash-based profitability, accrual quality, balance-sheet defensiveness), management decisions (net operating assets and R&D intensity), valuation level (free cash flow yield

and dividend yield), and market behavior and patterns (short-horizon reversal and long-lag seasonality). Posterior analysis of the market prices of risk further identifies `seas_6_10an`, `seas_11_15an`, `rd_me`, and `rmax5_rvol_21d` as the survivors whose 95% credible intervals exclude zero, marking them as the strongest candidates for inclusion in the true SDF beyond the benchmark menu.

The rest of the paper is organized as follows. Section 2 develops our Bayesian spanning framework, beginning with single anomalies and extending sequentially to pairs, triplets, and larger sets under EFDR control. Section 3 presents simulation evidence under baseline, benchmark-uncertainty, and masking designs. Section 4 reports the empirical analysis and shows that most anomalies in Jensen et al. (2023) are spanned, leaving a much smaller residual set. Section 5 provides post-selection validation and benchmark-robustness analysis. Section 6 concludes.

2 Methodology: Which Anomalies Can Be Spanned?

Our methodology develops in three stages, each addressing a limitation of the previous one. We begin with a Bayesian version of the conventional fixed-benchmark test, which improves on the frequentist approach by delivering posterior spanning probabilities and multiplicity control. We then generalize to allow for model uncertainty, integrating over all admissible partitions of the benchmark factors rather than conditioning on a single model. Finally, we extend to a sequential procedure that evaluates anomalies jointly in groups of increasing size, resolving masking and redundancy that single-anomaly tests cannot detect. Throughout, we use `FF6` as the starting benchmark, though the framework applies to any benchmark.

2.1 Bayesian Test with a Fixed Benchmark

Suppose we have n candidate anomalies $a_1, a_2, \ldots, a_n$. For each anomaly $i = 1, \ldots, n$, we test the pair of hypotheses

$$H_{0,i} : a_i \text{ is not spanned by the benchmark}, \tag{1}$$

$$H_{1,i} : a_i \text{ is spanned by the benchmark}. \tag{2}$$

The conventional approach tests the inverted version of these hypotheses using an alpha test against a fixed benchmark, but as explained in the introduction, this approach cannot establish spanning affirmatively, and its conclusions are benchmark-conditional. Our Bayesian approach addresses the first limitation by delivering a posterior probability of spanning that constitutes direct evidence in favor of our alternative. We will show how the second limitation is overcome below.

In the fixed-benchmark Bayesian test for anomaly a_i, we compare two models defined by different splits of the augmented factor set $\boldsymbol{f}_i = (\texttt{FF6}, a_i)$ into unspanned and spanned components. In the first model, we place all six `FF6` factors in the unspanned set and a_i in the spanned set,

$$\underbrace{\texttt{FF6}}_{=\boldsymbol{x}_{i,120}} = \boldsymbol{\lambda}_{i,120} + \boldsymbol{u}_{i,120}, \tag{3}$$

$$\underbrace{a_i}_{=\boldsymbol{w}_{i,120}} = \Gamma_{i,120}\, \boldsymbol{x}_{i,120} + \boldsymbol{\varepsilon}_{i,120}, \tag{4}$$

so that a_i lies in the span of `FF6`. In the second model, we treat a_i as an additional unspanned factor alongside `FF6`,

$$\underbrace{\begin{pmatrix} \texttt{FF6} \\ a_i \end{pmatrix}}_{=\boldsymbol{x}_{i,127}} = \boldsymbol{\lambda}_{i,127} + \boldsymbol{u}_{i,127}, \qquad \boldsymbol{w}_{i,127} = \varnothing, \tag{5}$$

so that a_i is not within the span of FF6. We use subscript i to indicate that this is the split involving anomaly a_i. The subscripts 120 and 127 refer to the model numbers in the full space of models considered below.

Under Gaussian errors and the prior of Chib and Zeng (2020), specialized in Chib, Zeng, and Zhao (2020), the marginal likelihood (Chib, 1995) of each model is available in closed form. The posterior spanning probability for anomaly i under this two-model comparison is

$$p_i^{\text{fixed}} = \Pr(a_i \text{ is spanned} \mid \text{data}) = \frac{m_{i,120}}{m_{i,120} + m_{i,127}}, \tag{6}$$

where $m_{i,120}$ and $m_{i,127}$ are the marginal likelihoods of the two models for anomaly i under equal prior weights. A value of p_i^{fixed} close to one constitutes affirmative evidence that a_i is spanned. This is something the frequentist test cannot provide.

EFDR control with a fixed benchmark. When n anomalies are tested simultaneously, multiplicity must be addressed. We adopt the *expected false discovery rate* (EFDR) procedure from the Bayesian multiple testing literature (Müller et al., 2007; Newton et al., 2004). Let p_i^{fixed} be the posterior spanning probability for anomaly i and define the local false discovery rate

$$\text{lfdr}_i = 1 - p_i^{\text{fixed}}. \tag{7}$$

If a rule selects a set $S \subset \{1, \ldots, n\}$ as spanned, the posterior expected false discovery rate is

$$\text{EFDR}(S \mid \text{data}) = \frac{\sum_{i \in S} \text{lfdr}_i}{\max\{|S|, 1\}}. \tag{8}$$

Sorting the lfdr_i in increasing order and choosing the largest k^* such that $\text{EFDR}(k^*) \leq q$ for a prespecified level $q \in (0, 1)$ yields the spanned set $S^* = \{a_{(1)}, \ldots, a_{(k^*)}\}$. Proposition 1 below formalizes the decision-theoretic optimality of this rule.

The fixed-benchmark Bayesian test with EFDR control is already an improvement over standard frequentist practice. It delivers posterior evidence for spanning rather

than merely failing to reject and controls the expected fraction of errors among declared discoveries. However, it inherits the second difficulty identified in the introduction: the conclusion depends on whether $\mathbb{M}_{i,120}$ in Equations (3) and (4) is the correct spanning model. If a_i is actually spanned by a proper subset of FF6, neither $\mathbb{M}_{i,120}$ nor $\mathbb{M}_{i,127}$, the model defined by Equations (5) and (6), captures this, and the two-model comparison can give misleading evidence in either direction. This motivates the next step.

2.2 Integrating Model Uncertainty

The two models $\mathbb{M}_{i,120}$ and $\mathbb{M}_{i,127}$ are just two of the $J = 2^7 - 1 = 127$ nontrivial ways to partition the augmented factor set $\boldsymbol{f}_i = (\texttt{FF6}, a_i)$ into unspanned and spanned components. Each such partition defines a model $\mathbb{M}_{i,j}$ of the form

$$\boldsymbol{x}_{i,j} = \boldsymbol{\lambda}_{i,j} + \boldsymbol{u}_{i,j}, \tag{9}$$

$$\boldsymbol{w}_{i,j} = \Gamma_{i,j}\boldsymbol{x}_{i,j} + \boldsymbol{\varepsilon}_{i,j}, \qquad j = 1, 2, \ldots, 127, \tag{10}$$

where there is no intercept in the second equation since $\boldsymbol{w}_{i,j}$ is by definition spanned by $\boldsymbol{x}_{i,j}$. In some partitions, the anomaly a_i is an element of $\boldsymbol{x}_{i,j}$ (unspanned); in others, it is an element of $\boldsymbol{w}_{i,j}$ (spanned). Table A.1 in the appendix enumerates all 127 partitions. The unshaded rows collect models $\mathcal{M}_0$ in which a_i is unspanned; the shaded rows collect models $\mathcal{M}_1$ in which a_i is spanned. There are $|\mathcal{M}_0| = 64$ and $|\mathcal{M}_1| = 63$ models in each class. Starting with a discrete uniform prior over the 127 models,

$$\Pr(\mathbb{M}_{i,j}) = \frac{1}{J}, \tag{11}$$

the prior probability of $H_{0,i}$ and $H_{1,i}$ in hypotheses (1) and (2) are essentially equal, so this prior does not favor either hypothesis. By Bayes' theorem, the posterior probability of each model is

$$\Pr(\mathbb{M}_{i,j} \mid \text{data}) = \frac{m_{i,j}}{\sum_{\ell=1}^{J} m_{i,\ell}}, \qquad j = 1, \ldots, J, \tag{12}$$

where $m_{i,j}$ is the marginal likelihood of $\mathbb{M}_{i,j}$, available in closed form under the prior of Chib and Zeng (2020). The posterior spanning probability for anomaly a_i, integrating over all models in which a_i is spanned, is

$$p_i = \Pr(a_i \text{ is spanned} \mid \text{data}) = \sum_{j:\, a_i \in \boldsymbol{w}_{i,j}} \Pr(\mathbb{M}_{i,j} \mid \text{data}). \tag{13}$$

This generalizes p_i^{fixed} from the previous section. It includes the evidence from $\mathbb{M}_{i,120}$ (full-benchmark spanning) and $\mathbb{M}_{i,127}$ (no spanning), but also from the 62 additional models where a_i is spanned by a proper subset of `FF6`. If a_i is spanned by, say, the market and value factors alone, the posterior weight will concentrate in those models, and the total spanning probability p_i will be large even though $\mathbb{M}_{i,120}$ might not dominate. The fixed-benchmark test misses this possibility entirely. Applying the EFDR selector to $p_1, \ldots, p_n$ computed under the full 127-model space yields a spanning set S^* that is robust to which subset of `FF6` happens to span each anomaly. The benchmark-dependence limitation identified in the introduction is resolved: the conclusion no longer requires the analyst to have specified the correct spanning model in advance.

We formally summarize the decision-theoretic properties of the singleton step with model uncertainty in the following proposition.

Proposition 1 (Bayes-optimal EFDR-controlled selection)**.** *Let*

$$k^* = \max\Big\{k \in \{0, \ldots, n\} : \frac{1}{k}\sum_{\ell=1}^{k} \text{lfdr}_{(\ell)} \leq q\Big\}$$

and set $S^* = \{(1), \ldots, (k^*)\}$*. Then*

1. $\text{EFDR}(S^* \mid \textit{data}) \leq q$*.*
2. *Among all* S *with* $\text{EFDR}(S \mid \textit{data}) \leq q$*,* S^* *maximizes the posterior expected number of true discoveries* $\sum_{i \in S} p_i$*.*

Proof. (1) By construction, $\text{EFDR}(S^* \mid \text{data})$ is the average of the k^* smallest lfdr values,

which is $\leq q$. (2) Fix $k = |S|$. Swapping any omitted $\text{lfdr}_{(j)}$ for a larger included lfdr weakly tightens the constraint and strictly increases $\sum_{i \in S}(1 - \text{lfdr}_i)$. The optimal S of size k is therefore $\{(1), \ldots, (k)\}$; choosing the largest feasible k yields S^*. □

What model uncertainty costs in practice. Table 4 in the empirical section documents the cost of ignoring model uncertainty by comparing our full procedure against a restricted two-model specification. At Step 1, ignoring model uncertainty classifies 58 anomalies as fake versus 21 under the full procedure. The difference of 37 anomalies represents factors that appear spanned under a specific benchmark model but carry independent pricing information once the full distribution of SDFs is integrated out. The inflated fake-anomaly count under the restricted specification is spurious: it reflects the constraint that the spanning model must be $\mathbb{M}_{120}$, not genuine evidence of redundancy.

2.3 Sequential Joint Testing

Both steps above evaluate each anomaly individually. This leaves unresolved a phenomenon that is both statistically and economically important: *masking*. Two anomalies may each appear weakly spanned when evaluated alone, yet be jointly spanned when considered together, because the common variation they share relative to the benchmark only becomes detectable when both are present. A related issue is *redundancy*: multiple proxies for the same latent risk factor generate inflated marginal evidence when tested separately; joint testing reveals them as a single spanned cluster. Neither masking nor redundancy is visible at the individual-anomaly stage, regardless of whether one uses a fixed benchmark or integrates over model uncertainty.

We address this by extending the procedure sequentially to groups of anomalies of increasing size. After applying the EFDR-controlled step-up rule to singleton posteriors, let $\mathcal{A}_2^*$ denote the anomalies not yet classified as spanned. We then consider all pairs of anomalies from $\mathcal{A}_2^*$. For each pair $\mathcal{G}_i = \{a_{i,1}, a_{i,2}\}$, we form the augmented factor set

$$\boldsymbol{f}_i = (\texttt{FF6}, a_{i,1}, a_{i,2}) \tag{14}$$

and enumerate the 127 nontrivial partitions of $\boldsymbol{f}_i$ into spanned and unspanned components, restricting to models where the pair jointly occupies either $\boldsymbol{x}_j$ or $\boldsymbol{w}_j$ (Table A.2). The joint posterior spanning probability is

$$p_i^{\text{joint}} = \sum_{j:\,\mathcal{G}_i \subseteq \boldsymbol{w}_j} \Pr(\mathbb{M}_{ij} \mid \text{data}), \tag{15}$$

and the local false discovery rate for the pair is $\text{lfdr}_i = 1 - p_i^{\text{joint}}$. We apply the same EFDR step-up rule to the $N_2 = \binom{|\mathcal{A}_2^*|}{2}$ pairs, declaring jointly spanned those with the smallest lfdr values subject to $\text{EFDR} \leq q$.

At generic stage s, let $\mathcal{A}_s^*$ denote the anomalies remaining from step $s-1$ and let $N_s = \binom{|\mathcal{A}_s^*|}{s}$ denote the number of candidate subsets of size s. For each subset $\mathcal{G}_i$, we test

$$H_{0,i} : \mathcal{G}_i \text{ is not jointly spanned by any benchmark,} \tag{16}$$

$$H_{1,i} : \mathcal{G}_i \text{ is jointly spanned by some benchmark.} \tag{17}$$

Mixed configurations are excluded: the anomalies in $\mathcal{G}_i$ either jointly appear in $\boldsymbol{w}_j$ or jointly appear in $\boldsymbol{x}_j$. After applying the EFDR step-up rule, the spanned set at stage s is

$$\mathcal{S}_s^* = \bigcup_{\ell=1}^{k_s^*} \mathcal{G}_{(\ell)}, \tag{18}$$

and the unspanned pool is updated as $\mathcal{A}_{s+1}^* = \mathcal{A}_s^* \setminus \mathcal{S}_s^*$. The process continues until no further anomalies are classified as spanned.

The decision-theoretic properties of our sequential procedure are summarized in the next result.

Proposition 2 (Sequential EFDR control across stages)**.** *Fix data $\mathcal{D}$ and $q \in (0,1)$. At each stage s, apply the step-up rule of Proposition 1 to the subset-level hypotheses $\mathcal{H}_s$, yielding $\mathcal{S}_s^* \subseteq \mathcal{H}_s$. Assume $\mathcal{H}_{s+1} \subseteq \mathcal{H}_s \setminus \mathcal{S}_s^*$, so selections $\mathcal{S}_1^*, \ldots, \mathcal{S}_S^*$ are pairwise disjoint. Then the*

cumulative discovery set $\mathcal{S}^* = \bigcup_{s=1}^{S} \mathcal{S}_s^*$ *satisfies*

$$\mathrm{EFDR}(\mathcal{S}^* \mid \mathcal{D}) \leq q.$$

Proof. By Proposition 1, $\sum_{\mathcal{S} \in \mathcal{S}_s^*} \mathrm{lfdr}_{\mathcal{S}} \leq q \cdot |\mathcal{S}_s^*|$ at each stage. Disjointness of selections gives $\sum_{\mathcal{S} \in \mathcal{S}^*} \mathrm{lfdr}_{\mathcal{S}} \leq q \sum_s |\mathcal{S}_s^*| = q|\mathcal{S}^*|$, so $\mathrm{EFDR}(\mathcal{S}^* \mid \mathcal{D}) \leq q$. If every $\mathcal{S}_s^* = \varnothing$, then $\mathrm{EFDR} = 0 \leq q$ by convention. □

Proposition 2 establishes that the sequential screening process preserves EFDR control globally, not just within each stage. Because discovered sets are removed from subsequent testing, the posterior expected proportion of false discoveries in the cumulative spanned set remains bounded by q at all stages simultaneously.

2.4 Robustness to Omitted SDF Components

Our two theoretical results deliver decision-theoretic guarantees conditional on the payoff space examined at each stage. A separate concern is whether spanning conclusions drawn from a restricted payoff universe could be overturned by components of the true SDF that lie outside that universe. The following proposition shows that this concern is unfounded.

Proposition 3 (Projection irrelevance of omitted SDF components)**.** *Let*

$$\mathcal{V}_s = \mathrm{span}\{f_1, \ldots, f_K\} \subset L^2$$

denote the payoff span at stage s*, and let* $m^\star \in L^2$ *be any SDF pricing all payoffs in* $\mathcal{V}_s$*. For every* $r \in \mathcal{V}_s$*,* $\mathbb{E}[m^\star r] = \mathbb{E}[P_{\mathcal{V}_s} m^\star \cdot r]$*, and any spanning statement formulated solely in terms of linear relations among payoffs in* $\mathcal{V}_s$ *is invariant to the component of* $m^\star$ *orthogonal to* $\mathcal{V}_s$*.*

Proof. Write $m^\star = m_{\mathcal{V}_s} + m_\perp$ with $m_\perp \perp \mathcal{V}_s$. For $r \in \mathcal{V}_s$, $\mathbb{E}[m^\star r] = \mathbb{E}[m_{\mathcal{V}_s} r] + \mathbb{E}[m_\perp r] = \mathbb{E}[m_{\mathcal{V}_s} r]$ since $m_\perp \perp r$. Spanning statements depend on the pricing kernel only through its action on $\mathcal{V}_s$, which is fully captured by $m_{\mathcal{V}_s}$. □

Any component of the true SDF orthogonal to the working payoff space $\mathcal{V}_s$ has no effect on pricing or spanning restrictions evaluated within $\mathcal{V}_s$. Omitted SDF directions therefore cannot overturn the classifications our procedure delivers. The procedure does not aim to identify the true SDF globally; it delivers internally valid spanning decisions conditional on the observable payoff space at each stage.

2.5 Why Spanning is the Right Burden of Proof

The three-step progression above rests on the statistical asymmetry described in the introduction, and it is worth making this asymmetry precise in light of the machinery just developed. At each stage, declaring an anomaly spanned requires only that some model $\mathbb{M}_j$ with $a \in \boldsymbol{w}_j$ receives substantial posterior weight. This is a directly verifiable statement about the observable payoff space. The monotonicity property, which states that an anomaly spanned by one benchmark is spanned by any richer benchmark containing it, ensures that a positive finding at any stage is conclusive, regardless of whether the benchmark is globally optimal. Declaring an anomaly unspanned, by contrast, would require ruling out all possible spanning models, including those in which a is spanned by benchmarks not yet considered. Our framework places the burden of proof on the easier object and treats the harder claim as the default that the data must overcome.

3 Simulation Study

We verify the finite-sample properties of the methodology developed in Section 2 in a Monte Carlo study where spanning status is known by construction. The three scenarios are introduced in the order in which the methodology was motivated. Scenario A evaluates baseline classification performance; Scenario B isolates the additional gain from integrating model uncertainty; and Scenario C isolates the gain from sequential joint testing.

3.1 Common Setup

We fix the benchmark to FF6, set $n = 30$ candidate anomalies and $T = 480$ months to match the empirical sample, and use 500 replications throughout with EFDR target $q = 0.20$. For a truly spanned anomaly i,

$$a_{it} = \Gamma_i^\top f_t + \varepsilon_{it}, \qquad \mathbb{E}[\varepsilon_{it}] = 0, \tag{19}$$

where f_t denotes the FF6 returns at time t. For a truly unspanned anomaly j,

$$a_{jt} = \alpha_j + \Gamma_j^\top f_t + \varepsilon_{jt}, \qquad \alpha_j \neq 0, \tag{20}$$

where the nonzero intercept captures the genuine pricing signal. Loadings Γ, residual variances, and the factor covariance matrix are calibrated to the observed FF6 and Jensen et al. (2023) data, and α is calibrated to target full-benchmark alpha t-statistics rather than raw intercept magnitudes. Each replication uses the same 25%/75% training-evaluation split as the empirical analysis.

We focus on two signal regimes throughout. In the *moderate* regime, truly unspanned anomalies have target alpha t-statistics of roughly 2 to 3. In the *weak* regime, the corresponding range is 0.5 to 1.0. These regimes mirror the empirical setting where signal-to-noise ratios are low and strong-signal cases are comparatively uninformative.

We compare Bayes EFDR against five procedures in every scenario. The first four are fixed-benchmark time-series alpha tests based on the full FF6 model: no multiplicity correction, Benjamini–Hochberg (BH), Bonferroni, and the Harvey et al. (2016) hurdle ($|t| > 3$). The fifth is a fixed-benchmark Bayes test that selects $\mathbb{M}_{120}$ (spanned) over $\mathbb{M}_{127}$ (unspanned) if the posterior probability of spanning exceeds 0.75 and the latter if it is less than 0.25, leaving an undecided region between the two thresholds. All methods are evaluated on identical simulated panels using common random seeds.

We calculate the following metrics. Let S^* and U^* denote the sets declared spanned and unspanned. We report

$$\text{True Positive Rate (TPR)} = \frac{|\{i \in S^* \text{ and } i \text{ truly spanned}\}|}{n_1}, \tag{21}$$

$$\text{True Negative Rate (TNR)} = \frac{|\{i \in U^* \text{ and } i \text{ truly unspanned}\}|}{n_2}, \tag{22}$$

$$\text{False Discovery Rate (FDR)} = \frac{|\{i \in S^* \text{ and } i \text{ truly unspanned}\}|}{\max(|S^*|, 1)}, \tag{23}$$

$$\text{False Non-Discovery Rate (FNDR)} = \frac{|\{i \in U^* \text{ and } i \text{ truly spanned}\}|}{\max(|U^*|, 1)}. \tag{24}$$

All four metrics are averaged across replications. For Bayes testing, undecided anomalies are counted in neither S^* nor U^*. In addition, it is worth noting that the Bayes EFDR procedure controls the posterior expected FDR within a given replication at level $q = 0.20$. This is the quantity defined in Proposition 1 and is a statement about the posterior distribution given the data in hand. The FDR reported in the simulation tables is the average across replications of the realized FDR in each replication, which is a frequentist sampling average over draws of the data. These two objects are related but not identical, and there is no theoretical claim that the simulation-average FDR should be bounded by q. In the weak-signal regime, the simulation-averaged FDR is larger because borderline anomalies are more common, and the EFDR selector includes more of them to maximize the number of true discoveries subject to the posterior constraint. The FDR column should therefore be read as a measure of the procedure's precision in the frequentist sense, not as a test of whether the theoretical guarantee is satisfied.

The frequentist procedures occupy the first four columns of each table. Because nonrejection of $H_0 \colon \alpha = 0$ is not affirmative evidence of spanning, TPR and FDR are not defined for these methods and are shown as dashes. The dashes are not missing entries; they reflect the structural inability of rejection-based tests to make positive spanning classifications, which is the limitation that motivates our approach. TNR and FNDR are defined for the frequentist procedures and appear in the usual way. In

the conventional frequentist framing, where the null is that the anomaly is spanned and rejection declares it unspanned, FNDR corresponds to the Type I error rate and FDR to the Type II error rate. The frequentist procedures control FNDR at the 5% level by construction, which is why those entries are small. FDR is not controlled and can be large when the signal is weak, reflecting low power to detect truly unspanned anomalies. Bayes EFDR, by contrast, maximizes TPR subject to a bound on the posterior expected FDR via Proposition 1, which is why it achieves substantially higher TPR at the cost of a higher FNDR relative to the frequentist procedures. TPR is the primary metric of interest for our method since the goal is to identify as many truly spanned anomalies as possible while maintaining meaningful discrimination on the unspanned side.

3.2 Scenario A: Baseline Classification

This design evaluates the baseline performance of the Bayesian procedure when no masking or model uncertainty is present. We generate 15 truly spanned anomalies as full-$\texttt{FF}\,6$ linear combinations with zero intercepts, and 15 truly unspanned anomalies with nonzero intercepts calibrated to the target signal regime. This is the cleanest possible environment: if a method fails to distinguish the two types here, it has little chance in the data.

Bayes EFDR delivers substantially higher TPR than all competing procedures in both signal regimes: 0.883 versus 0.566 for Bayes testing in the moderate regime, and 0.899 versus 0.566 in the weak regime. The fixed-benchmark Bayes test is highly conservative: its symmetric $\pm \log 3$ rule creates a large undecided region that suppresses affirmative spanning classifications at a severe cost to power. The frequentist procedures cannot be assessed on the spanned side at all; their reported TNR values reflect the behavior of a test designed to reject rather than confirm. Bayes EFDR maintains a TNR of 0.625 (moderate) and 0.150 (weak), showing that the higher TPR does not come from indiscriminate classification. The TNR decline in the weak regime is expected: when

Table 1 Simulation Evidence: Scenario A (Baseline Classification).

Note: 15 truly spanned and 15 truly unspanned anomalies, $n = 30$, $T = 480$, 500 replications, $q = 0.20$. Boldface marks the row-wise best displayed value. Higher is better for TPR and TNR; lower is better for FDR and FNDR. Dashes indicate that the metric is not defined for frequentist procedures (see text).

Regime	Metric	No corr.	BH	Bonferroni	HLZ	Bayes testing	Bayes EFDR
Moderate	TPR	–	–	–	–	0.566	**0.883**
	TNR	**0.634**	0.424	0.214	0.265	0.417	0.625
	FDR	–	–	–	–	**0.170**	0.290
	FNDR	0.074	0.025	**0.006**	0.008	0.052	0.147
Weak	TPR	–	–	–	–	0.566	**0.899**
	TNR	0.106	0.007	0.007	0.010	0.046	**0.150**
	FDR	–	–	–	–	**0.459**	0.485
	FNDR	0.316	**0.021**	**0.021**	0.035	0.232	0.372

signal-to-noise ratios are low, some genuinely unspanned anomalies have posteriors that approach the EFDR threshold, but the procedure correctly prioritizes recovery of the spanned set.

3.3 Scenario B: Gains from Model Uncertainty

This design isolates the specific contribution of integrating model uncertainty relative to a fixed benchmark. We generate 30 anomalies: 10 truly unspanned, 10 standard anomalies spanned by the full `FF6` benchmark (type i), and 10 anomalies spanned only by a proper subset of the benchmark, taken in the baseline implementation to be (`Mkt`, `HML`) (type ii). Type-ii anomalies have true zero intercepts relative to the correct subset model but nonzero intercepts relative to the full `FF6` model, so procedures that condition mechanically on the full benchmark will misclassify them as unspanned. Successful classification requires not merely detecting redundancy, but also identifying the correct benchmark partition.

The key comparison is between Bayes EFDR and fixed-benchmark Bayes testing on the type-ii anomalies, since these are the cases where model uncertainty matters

Table 2 Simulation Evidence: Scenario B (Model Uncertainty).

Note: 10 unspanned, 10 type-i spanned (full FF6), and 10 type-ii spanned (proper subset only), $n = 30$, $T = 480$, 500 replications, $q = 0.20$. TPR(type i) and TPR(type ii) report recovery rates within each spanned subgroup. Dashes indicate that the metric is not defined for frequentist procedures (see text).

Regime	Metric	No corr.	BH	Bonferroni	HLZ	Bayes testing	Bayes EFDR
Moderate	TPR (type i)	–	–	–	–	0.571	**0.889**
	TPR (type ii)	–	–	–	–	0.567	**0.901**
	TNR	**0.643**	0.375	0.214	0.267	0.427	0.619
	FDR	–	–	–	–	**0.096**	0.171
	FNDR	0.130	0.033	**0.012**	0.016	0.086	0.230
Weak	TPR (type i)	–	–	–	–	0.571	**0.901**
	TPR (type ii)	–	–	–	–	0.567	**0.916**
	TNR	0.114	0.009	0.007	0.012	0.057	**0.156**
	FDR	–	–	–	–	**0.296**	0.316
	FNDR	0.429	0.033	**0.030**	0.050	0.273	0.488

by design. In the moderate regime, Bayes EFDR achieves $\mathrm{TPR}(\text{type ii}) = 0.901$ against 0.567 for Bayes testing. In the weak regime the numbers are 0.916 and 0.567. The fixed-benchmark Bayes test performs nearly identically on type-i and type-ii anomalies (0.571 and 0.567 in both regimes), confirming that it has no mechanism to exploit the partition structure of the benchmark. Bayes EFDR recovers type-ii anomalies at least as well as type-i anomalies in both regimes, because the 127-model scan explicitly includes models where spanning is by a proper subset, and the posterior concentrates there when that is the correct partition. A procedure that conditions on a single model confuses "not spanned by this model" with "not spanned at all"; the simulation confirms that integrating model uncertainty resolves this confusion directly.

3.4 Scenario C: Gains from Sequential Joint Testing

This design isolates the gain from sequential joint testing over evaluating anomalies one at a time. The 30 anomalies comprise 28 standard anomalies (13 truly spanned and 15 truly unspanned) and two *masked* anomalies. The masked anomalies are truly spanned by construction, but generated from a low-dimensional benchmark subset,

by default, the market and value factors with a common loading direction and high idiosyncratic noise variance. Individually, this noise makes their full-benchmark alpha estimates too imprecise for reliable classification: each appears only weakly spanned when examined alone. Jointly, however, they span the same subspace, so Step 3 should recover them together once the common variation can be pooled. This design isolates the economic mechanism of masking rather than treating it as a mechanical multiple-testing artifact.

In addition to TPR, TNR, FDR, and FNDR computed over all 30 anomalies, we report two masking-specific diagnostics. "Masked share spanned" is the average fraction of the two masked anomalies declared spanned per replication. "All masked spanned" is the fraction of replications in which both masked anomalies are simultaneously declared spanned.

Table 3 Simulation Evidence: Scenario C (Masking).

Note: 13 standard spanned, 15 unspanned, and 2 masked anomalies, $n = 30$, $T = 480$, 500 replications, $q = 0.20$. "Masked share spanned" is the average fraction of the two masked anomalies declared spanned per replication. "All masked spanned" is the fraction of replications in which both are simultaneously declared spanned. Dashes indicate that the metric is not defined for frequentist procedures (see text).

Regime	Metric	No corr.	BH	Bonferroni	HLZ	Bayes testing	Bayes EFDR
Moderate	TPR	–	–	–	–	0.574	**0.879**
	TNR	**0.635**	0.435	0.218	0.269	0.424	0.627
	FDR	–	–	–	–	**0.171**	0.289
	FNDR	0.074	0.029	**0.006**	0.012	0.054	0.149
	Masked share spanned	–	–	–	–	0.571	**0.906**
	All masked spanned	–	–	–	–	0.336	**0.820**
Weak	TPR	–	–	–	–	0.574	**0.898**
	TNR	0.110	0.009	0.008	0.012	0.049	**0.148**
	FDR	–	–	–	–	**0.458**	0.486
	FNDR	0.313	0.025	**0.022**	0.043	0.222	0.385
	Masked share spanned	–	–	–	–	0.571	**0.909**
	All masked spanned	–	–	–	–	0.336	**0.828**

The masking diagnostics tell the clearest story. Bayes EFDR declares 90.6% of masked anomalies spanned in the moderate regime and 90.9% in the weak regime; Bayes testing reaches only 57.1% in both. The gap is larger still for complete recovery:

the fraction of replications in which both masked anomalies are simultaneously declared spanned is 82.0% and 82.8% under Bayes EFDR, against 33.6% under Bayes testing in both regimes. Additional stage-level diagnostics confirm the mechanism: at least one masked anomaly is first recovered at Step 2 in roughly 73% of replications in both regimes, and both are first recovered simultaneously at Step 2 in 30.0% of moderate-regime replications and 28.6% of weak-regime replications. No masking recovery occurs at Step 1 in any replication, confirming that the sequential structure is doing the work rather than the individual-anomaly scan.

Reading the three tables together, the overall picture is clear. The frequentist procedures cannot be assessed on the spanned side and provide at best indirect evidence of redundancy through rejection. The fixed-benchmark Bayes test improves on them by delivering posterior spanning probabilities but is too conservative and blind to both model uncertainty and masking. Bayes EFDR dominates on every margin that requires a positive finding: recovery of truly spanned anomalies, resolution of masking, and classification of anomalies spanned only by a proper benchmark subset. It is the procedure we carry to the data.

4 Empirical Study

We obtain U.S. monthly data for the benchmark factors `Mkt`, `SMB`, `HML`, `RMW`, `CMA`, and `MOM` from the Fama–French data library. The anomaly universe is the 153-factor database of Jensen, Kelly, and Pedersen (2023), grouped into 13 thematic categories. We use monthly value-weighted returns from January 1985 through December 2024, yielding $T = 480$ months. For the Bayesian exercises, the first 25% of the sample ($T_{\text{tr}} = 120$) is used to calibrate priors and the remaining $T_{\text{est}} = 360$ observations are used for model estimation.[1] $\boldsymbol{w}$ denotes the set affirmatively classified as *spanned* by the procedure, which we refer to as *fake anomalies*, and $\boldsymbol{x}$ denotes the *survivor set*. The interpretation is asymmetric: membership in $\boldsymbol{w}$ is a positive spanning classification,

[1] In the non-Bayesian diagnostics, we continue to use the full sample of $T = 480$ months.

confirming that the anomaly is *fake*, whereas membership in $\boldsymbol{x}$ should not be read as proof that the anomaly is genuinely unspanned or belongs to the true SDF. It means only that the anomaly survives the screening procedure and is not affirmatively classified as spanned relative to the benchmark and the candidate anomaly pool.

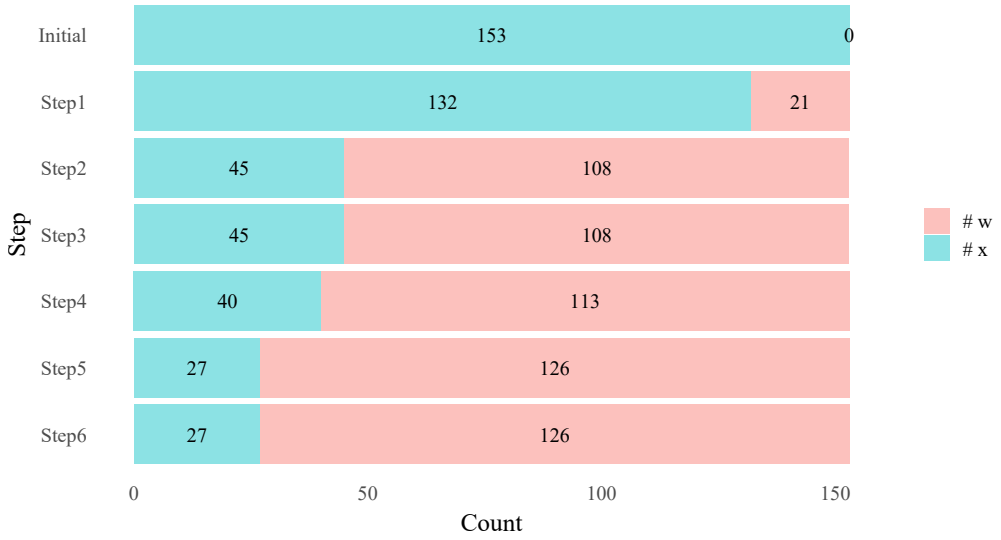
4.1 Headline Classification under FF6

We begin with the main empirical question: relative to FF6, how many of the 153 proposed anomalies can be affirmatively classified as spanned once benchmark subsets and joint spanning are allowed? To answer this, we apply the sequential Bayes–EFDR procedure with $q = 0.20$ to the full anomaly universe. At Step s, the benchmark is evaluated jointly with subsets of s candidate anomalies. Once an anomaly or group is classified as spanned, its members are deemed as *fake anomalies*, moved to $\boldsymbol{w}$, and removed from later stages; the remaining anomalies form the updated survivor pool.

Figure 2 summarizes the identification path. Steps 1, 2, 4, and 5 contribute 21, 87, 5, and 13 newly classified spanned anomalies, respectively, while Step 3 contributes none. The procedure stabilizes at 126 anomalies in $\boldsymbol{w}$ and 27 anomalies in $\boldsymbol{x}$, implying that 82.4% of the proposed anomaly zoo can be affirmatively classified as spanned relative to FF6 and the candidate anomaly menu explored by the procedure. The evidence is strong on the spanned side: 126 anomalies can be verified as redundant. The 27 survivors are the anomalies that remain after this screening, not a set proven to coincide with the true SDF. This suggests that the vast majority of proposed factors provide no incremental pricing information after rigorous spanning evaluation. Only a relatively small subset of anomalies remains capable of contributing meaningfully to the SDF, highlighting substantial redundancy within the anomaly zoo.

The identification path is also highly concentrated. Of the 126 anomalies eventually placed in $\boldsymbol{w}$, 108 are identified by the first two stages alone, and Step 2 by itself contributes 87. This concentration is already suggestive of masking: many anomalies that do not provide enough evidence of spanning in isolation become classifiable once

Figure 2 Counts of Fake Anomalies and Survivors.
Note: This figure reports the counts of anomalies classified into the spanned set $\boldsymbol{w}$ and the survivor set $\boldsymbol{x}$ when `FF6` is the benchmark. Green bars denote survivors ($\boldsymbol{x}$), and red bars denote anomalies classified as fake ($\boldsymbol{w}$).

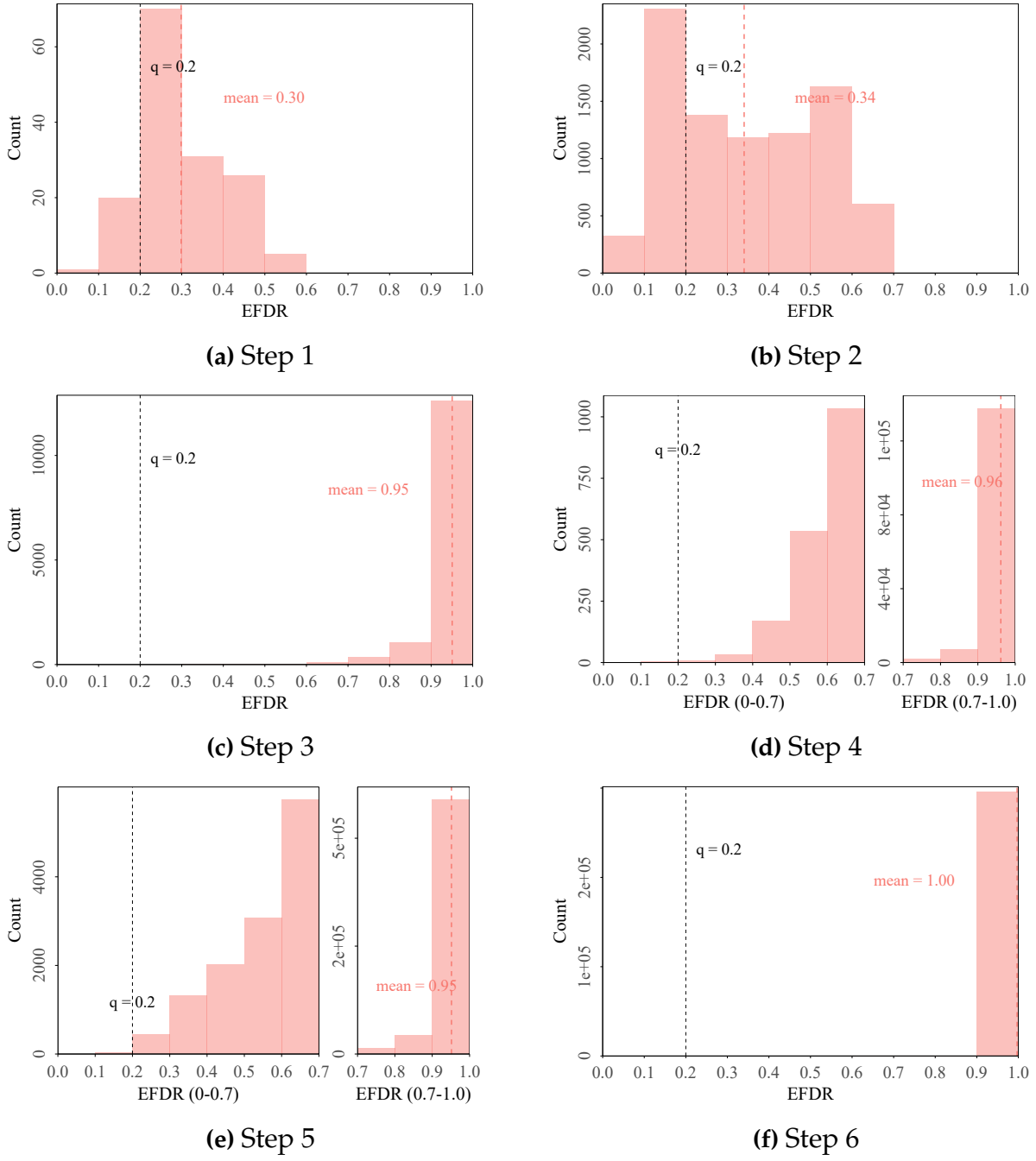


joint redundancy is allowed to emerge.

Robustness to the EFDR threshold The headline split is not especially sensitive to the choice of the EFDR threshold. Figures A.2 and A.3 show that when q varies between 0.10 and 0.20, the final count of spanned anomalies is stable, with the main differences appearing in the timing of when classifications occur rather than in the eventual classification itself. Meanwhile, it is essential to note that $q = 0.2$ is not a strict threshold. Figure A.4 displays the bar plot of lfdr and EFDR in Step 1. Even though the maximum EFDR in this step is about 0.5, choosing $q = 0.2$ corresponds to a relatively low quantile, meaning that the factors classified as $\boldsymbol{w}$ in step 1 are screened under a fairly strict standard.

Figure 3 provides a complementary view by plotting the distribution of candidate-set EFDR values at each stage. Statistically, the absence of new $\boldsymbol{w}$ in Step 3 reflects a selection-exhaustion effect: Steps 1 and 2 have already removed the anomalies most easily spanned. In Steps 1 and 2, EFDR values fall primarily between 0.1 and 0.6, so

Figure 3 Histogram of EFDRs across Steps 1–6.
Note: This figure displays, for steps 1–6 under the `FF6` benchmark, overlaid histograms of EFDR (red). We also mark the threshold at $q = 0.2$ and report the mean EFDR in each panel.



(a) Step 1 **(b)** Step 2

(c) Step 3 **(d)** Step 4

(e) Step 5 **(f)** Step 6

adopting $q = 0.2$ as the threshold identifies a subset of $\boldsymbol{w}$ factors. From Step 3 onward, the distribution shifts sharply toward one, indicating that most remaining candidate groups provide little posterior support for a spanning classification. As the set of explained $\boldsymbol{w}$ expands and the candidate pool contracts, locating additional $\boldsymbol{w}$ among the remaining anomalies becomes increasingly difficult, accounting for the sharp drop in incremental detections at later steps.

4.2 Why Model Uncertainty Matters

The sequential results above integrate over the full model space of admissible benchmark partitions. To isolate the role of considering model uncertainty, we conduct a counterfactual experiment. We compare our framework against a restricted specification that relies on only two models ($\mathbb{M}_{i,120}$ and $\mathbb{M}_{i,127}$) for each augmented set $\boldsymbol{f}_i$. Table 4 reports the cumulative number of anomalies classified as spanned by each stage under alternative restrictions on the model space.

The results reveal that ignoring model uncertainty leads to a substantial overestimation of the benchmark's explanatory power, resulting in "spurious spanning." This is most evident in the first step (Row 1). The restricted specification (Column (1)) aggressively classifies 58 anomalies as spanned. In contrast, our specification (Column (7)), which bases redundancy from the full distribution of SDFs, identifies only 21 spanned anomalies. By ignoring model uncertainty, the restricted approach prematurely discards potentially valid signals. Once model uncertainty is admitted, the evidence is no longer strong enough to classify those 37 anomalies affirmatively as spanned.

The same point reappears at later stages. Columns (2) through (6) represent hybrid specifications where the procedure begins with the robust baseline but reverts to the restricted set in a subsequent step. For instance, in Column (2), while the procedure correctly identifies 21 spanned anomalies in Step 1, the suppression of model uncertainty in Step 2 causes the count of spanned anomalies to surge to 129, significantly

Table 4 Cost of Ignoring Model Uncertainty

Note: This table compares the number of anomalies classified as spanned ($\boldsymbol{w}$) at each stage under two different model space specifications. Columns (1)-(6) report restricted specifications that ignore model uncertainty and rely on only two models. The shaded cells indicate the stage at which model uncertainty is suppressed. Column (1) suppresses model uncertainty throughout. Column (2) suppresses it only when the procedure reaches Step 2, and the remaining columns follow the same logic. Column (7) reports the baseline specification, which integrates over the full model space of 127 admissible partitions.

	Ignore Model Uncertainty						Consider Model Uncertainty
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Step1	58	21	21	21	21	21	21
Step2	116	129	108	108	108	108	108
Step3	116		117	108	108	108	108
Step4	127			132	113	113	113
Step5	127				142	126	126
Step6	127					135	126

exceeding the baseline count of 108. Each time that switch occurs, the cumulative count of anomalies assigned to $\boldsymbol{w}$ jumps discretely relative to the baseline. It confirms that the lower number of spanned anomalies in our baseline is not due to low statistical power, but rather due to the rigorous standard imposed by model averaging, which protects against the false discovery of redundancy.

4.3 Identification Path: Masking and Redundancy

Figure 2 unveils a striking pattern. A significant proportion of the fake anomalies ($\boldsymbol{w}$) are identified at the second stage, rather than at the first step. As discussed in Section 2.3, when a cluster of anomalies is redundant in the return space, essentially serving as alternative proxies for the same latent component, univariate testing tends to underestimate the evidence for spanning. Intuitively, while such proxies may appear to carry distinct pricing information when examined in isolation, their mutual substitutability is unveiled within a joint spanning framework. Thus, the second (and later) steps are crucial for identifying which anomalies are spanned.

Table 5 makes the masking channel more concrete by reporting, for each theme, the fraction of anomalies classified as spanned at each step. Several categories show little or no evidence of spanning in Step 1 and then a substantial reclassification in Step 2. Debt issuance, Low leverage, Size, and Value are the clearest examples: their Step-1 spanning shares are zero, but their Step-2 shares jump to 57.1%, 54.5%, 80.0%, and 72.2%, respectively. The natural reading is not that these themes become unimportant only at Step 2. It is that, for clustered signals, the one-at-a-time screen understates spanning evidence. Joint testing reveals that many of these variables are close substitutes for each other once the benchmark and the candidate anomalies are allowed to span together.

Table 5 Thematic Distribution of Fake Anomalies across Steps.

Note: This table reports, by economic theme, the percentage of anomalies that are classified as spanned at each stage of the sequential procedure. The first column lists the thematic categories following the classification in Jensen et al. (2023). The second column reports the total number of anomalies within each theme. Columns 3 through 7 display the percentage of anomalies in each theme that are identified as spanned in Steps 1 through 5 of the sequential testing procedure, respectively. Rows in bold highlight themes (Debt Issuance, Low Leverage, Size, and Value) that exhibit no spanned anomalies in Step 1 but a substantial proportion in Step 2, consistent with the masking effect.

Theme	# Factor in Each Theme	w (%) in Each Theme				
		Step 1	Step 2	Step 3	Step 4	Step 5
Accruals	6	16.7	33.3	0.0	0.0	0.0
Debt issuance	7	**0.0**	**57.1**	0.0	0.0	28.6
Investment	22	9.1	59.1	0.0	18.2	13.6
Low leverage	11	**0.0**	**54.5**	0.0	0.0	9.1
Low risk	18	22.2	77.8	0.0	0.0	0.0
Momentum	8	25.0	75.0	0.0	0.0	0.0
Profit growth	12	25.0	66.7	0.0	0.0	0.0
Profitability	11	9.1	45.5	0.0	0.0	9.1
Quality	17	5.9	23.5	0.0	0.0	29.4
Seasonality	12	33.3	50.0	0.0	0.0	0.0
Short term reversal	6	50.0	33.3	0.0	0.0	0.0
Size	5	**0.0**	**80.0**	0.0	0.0	0.0
Value	18	**0.0**	**72.2**	0.0	5.6	5.6

The dimensionality diagnostics show that the large Step-2 increment reflects redundancy rather than the removal of many independent pricing directions. Specifically,

to measure redundancy, we use the eigenvalues of the covariance matrices. Figure A.5 depicts the scree plots, where the steep decline in the eigenvalues of the fake anomalies identified in Step 2 shows that these factors share a sparse set of common drivers. Recognizing that set size can confound direct eigenvalue comparisons, we introduce the efficiency ratio as a size-adjusted metric. As reported in Table 6, the Efficiency Ratio for the w_2 set is notably low. Therefore, although Step 2 classifies a large number of anomalies as spanned, it discards very little independent pricing information. Instead, the procedure precisely purges dimensions that are highly collinear and statistically redundant.

Table 6 Effective Dimensionality of the Spanned and Survivor Sets.

Note: This table reports the dimensionality statistics for anomalies classified as spanned ($\boldsymbol{w}$) and those remaining in the survivor set ($\boldsymbol{x}$) at each stage of the stepwise procedure. Results for Step 3 are omitted, as no anomalies were classified as spanned at this stage. Panel A focuses on the specific set of anomalies identified as spanned in each step. Panel B analyzes the set of anomalies surviving after each step. "Number of Anomalies" denotes the count of anomalies in each set (N). $\text{Eff}_{\boldsymbol{N}}$ is the effective number of dimensions, calculated as $(\sum \lambda_i)^2 / \sum \lambda_i^2$, where λ_i are the eigenvalues of the covariance matrix of the factors in the set. "Eff Ratio" is the efficiency ratio, defined as Eff_N/N. A lower efficiency ratio indicates a higher degree of redundancy within the set.

Group of Anomalies	Number of Anomalies	Eff_N	Eff Ratio
Panel A. Anomalies classified as spanned at each step			
$\boldsymbol{w}$ in Step 1	21	4.64	0.22
$\boldsymbol{w}$ in Step 2	87	3.35	0.04
$\boldsymbol{w}$ in Step 4	5	1.73	0.35
$\boldsymbol{w}$ in Step 5	13	2.41	0.19
Panel B. Survivor set after each step			
Remaining $\boldsymbol{x}$ after Step 1	132	3.82	0.03
Remaining $\boldsymbol{x}$ after Step 2	45	3.90	0.09
Remaining $\boldsymbol{x}$ after Step 4	40	3.77	0.09
Remaining $\boldsymbol{x}$ after Step 5	27	4.11	0.15

4.4 What Characterizes the Survivor Set?

This subsection is descriptive. The objective is not to prove structural interpretations for the survivors, but to ask whether the pattern of classifications is economically coherent relative to the benchmark. Table 7 orders the 13 themes by the number of surviving anomalies. Once the FF6 dimensions are placed in the span, themes that are economically closest to those benchmark legs migrate heavily into $\boldsymbol{w}$. The concentration is stark for Investment, Momentum, and Low risk, each of which is fully classified as spanned. Investment signals are naturally close to CMA; momentum variants are close to MOM; and low-risk constructions are often reproduced by combinations of market, value, profitability, and investment exposures. This is the pattern one would expect if the procedure is isolating overlap rather than mechanically shrinking the anomaly zoo. Once FF6 is placed in the span, large parts of these themes are already replicated by benchmark exposures or by nearby anomalies that load on similar pricing channels.

Themes adjacent to other FF6 legs also largely migrate into $\boldsymbol{w}$. Profit growth (91.7% in $\boldsymbol{w}$) is closely tied to the earnings-based profitability leg RMW; Debt issuance (85.7%) comoves with investment and value through dilution and external-financing cycles, allowing CMA and HML to absorb much of the variation. Value (83.3%) and Size (80.0%) are primarily absorbed by HML and SMB, respectively. Seasonality (83.3%) captures predictable short-horizon return patterns driven by calendar effects and institutional rebalancing, which makes it closely related to momentum (MOM). Short-term reversal (80%) captures microstructure- and liquidity-driven short-horizon price reversals, which makes it negatively aligned with MOM and often accompanied by mild exposure to size (SMB). In such case, the economic mechanism behind the characteristic is already encoded in at least one FF6 factor, so a no-intercept spanning relation is easy to establish.

By contrast, themes that only partially overlap with FF6 leave more survivors in $\boldsymbol{x}$. Quality (58.8% in $\boldsymbol{w}$) and Profitability (63.6%) capture cash-flow durability

Table 7 Number and Percentage of Spanned and Surviving Anomalies by Theme.

Note: This table reports, by theme, the number and percentage of anomalies classified as spanned and the number and percentage that remain in the survivor set. Themes are ordered by the number of survivors in ascending order.

Theme	**#w**	**#x**	**(%)w**	**(%)x**
Investment	22	0	100.0%	0.0%
Low risk	18	0	100.0%	0.0%
Momentum	8	0	100.0%	0.0%
Profit growth	11	1	91.7%	8.3%
Debt issuance	6	1	85.7%	14.3%
Short-term reversal	5	1	83.3%	16.7%
Size	4	1	80.0%	20.0%
Seasonality	10	2	83.3%	16.7%
Value	15	3	83.3%	16.7%
Accruals	3	3	50.0%	50.0%
Low leverage	7	4	63.6%	36.4%
Profitability	7	4	63.6%	36.4%
Quality	10	7	58.8%	41.2%
Total	126	27	82.4%	17.6%

and operating efficiency, which RMW, built on operating profitability rather than cash, does not fully span. Accruals (50.0%) and related working-capital/NOA constructs primarily capture accounting recognition/timing effects and the composition of operating investment. These are only partially related to the "quantity of investment" in CMA. Low leverage (63.6%) primarily reflects financing frictions in the price and quantity of external capital, while also embedding profitability/safety, cash buffers, collateral/intangibles, and trading liquidity. Consequently, it is only imperfectly proxied by HML, SMB, or CMA. The result is a lower $\boldsymbol{w}$-share and a nontrivial $\boldsymbol{x}$-residual along these dimensions.

Overall, the closer a theme lies to the benchmark menu, the more likely it is to migrate into the spanned set. By contrast, survivors that deliver genuinely incremental pricing power tend to concentrate in more heterogeneous economic channels relative to the benchmark. Section 5 then asks whether the survivor set also differs from fake anomalies in out-of-sample predictive, portfolio, and pricing exercises.

5 Post-Selection Validation and Benchmark Robustness

Having established the headline screening results under `FF6`, we now ask whether the resulting classification is economically meaningful and robust. We proceed in four steps. First, we compare survivors and fake anomalies under the `FF6` benchmark from different perspectives. Second, we examine the sensitivity of the screening outcome to benchmark choice by rerunning the procedure under an alternative six-factor benchmark, `NewB6`. Third, we focus on the anomalies that survive under both benchmark menus and summarize the economic channels they represent. Finally, we evaluate the out-of-sample performance of this survivor set and use posterior market prices of risk to isolate the strongest candidates for augmenting the SDF.

5.1 Post-Selection Validation under FF6

Section 4.1 classifies 126 of the 153 candidate anomalies as spanned and retains 27 as survivors. We refer to the spanned group as S^* and the survivor group as U^* throughout this section. Because the screening rule is asymmetric, membership in S^* is an affirmative spanning classification, whereas membership in U^* means only that the anomaly was not found to be spanned. The evidence below should therefore be read as post-selection validation rather than as proof that U^* coincides with the true SDF. As in gold panning, the procedure filters out the sand; what remains contains gold but also some sand. If the classification is informative, sets drawn from U^* should outperform equally sized sets drawn from S^* in (i) predictive likelihood, (ii) portfolio efficiency, and (iii) cross-sectional fit. Since the survivor count stabilizes after the fifth step of the sequential procedure, we focus on five-factor blocks throughout.

For predictive likelihood, we randomly sample 1,000 distinct five-factor sets from U^* and, independently, 1,000 distinct five-factor sets from S^*, yielding 1,000,000 matched pairs. For each matched pair, we form a common factor universe consisting of the ten factors and estimate two specifications that differ only in the role assigned

to each block. In the survivor specification, the five factors drawn from U^* occupy the risk factor set $\boldsymbol{x}$ and the five drawn from S^* occupy the non-risk factor set $\boldsymbol{w}$. In the spanned specification, the roles are reversed: the five factors from S^* are placed in $\boldsymbol{x}$ and those from U^* in $\boldsymbol{w}$. The two specifications therefore use identical data but differ only in which block is treated as the source of priced risk. If the original classification is correct, the survivor specification should deliver higher out-of-sample predictive likelihood, since the factors in U^* carry genuine pricing information that the factors in S^* do not. We then repeat the same comparison after augmenting each block with `FF6`. The implementation follows Chib et al. (2024).[2] As for portfolio efficiency and cross-sectional fit, we compare Sharpe ratios and cross-sectional pricing performance across the same matched pairs, using the Fama–French 49 industry portfolios, 100 P-Tree portfolios (Cong et al., 2025), and 120 ($2 \times 3 \times 20$) bi-sort portfolios constructed from 20 characteristics.[3]

Table 8 reports the results. In both Panel A, which compares the two specifications using the ten sampled factors alone, and Panel B, which augments each specification with `FF6`, the survivor specification delivers higher out-of-sample predictive likelihood than the spanned specification across every reported statistic. In Panel A, the mean log average predictive likelihood under the survivor specification exceeds that under the spanned specification by approximately five units, and the same ordering holds after augmentation with `FF6`. Since the two specifications condition on the same ten factors and differ only in which block plays the role of $\boldsymbol{x}$, this difference is attributable entirely to the information content of the classification. The anomalies that survive the spanning screen carry more out-of-sample predictive information than equally sized sets drawn from the spanned group, which supports the validity of the original classifications.

We next turn from predictive likelihood to portfolio efficiency. For each matched

[2]We partition the 40-year sample into 30 years for estimation and 10 years for out-of-sample prediction, with the first 5 years used to fix the prior. We consider 5 contiguous estimation and prediction splits, re-estimating the model at each split and computing the predictive likelihood. The performance metric is the average across the 5 splits, and the computation is repeated for every factor combination.

[3]The characteristics are listed in Table A.3.

Table 8 Statistics of Log Average of Predictive Likelihoods for Different Factor Combinations.

Note: This table reports summary statistics for the log average predictive likelihood from matched factor-set comparisons. We randomly sample 1,000 five-factor sets from the survivor pool U^* and 1,000 five-factor sets from the spanned pool S^*, yielding 1,000,000 matched pairs. In each pair, one specification assigns the five factors from U^* to the survivor set $\boldsymbol{x}$ and the five from S^* to the spanned set $\boldsymbol{w}$; the other reverses these roles. Panel B repeats the same comparison after augmenting the full ten-factor universe with `FF6`.

Specification	Mean	Std.	Median	2.5% qtile	97.5% qtile
Panel A. Anomalies only					
$(x_1, \ldots, x_5) \in U^*,\ (w_1, \ldots, w_5) \in S^*$	**3241.5**	113.5	**3236.5**	**3033.8**	**3474.5**
$(x_1, \ldots, x_5) \in S^*,\ (w_1, \ldots, w_5) \in U^*$	3236.7	113.8	3231.8	3028.1	3470.1
Panel B. Anomalies augmented with `FF6`					
$(x_1, \ldots, x_5) \in U^*,\ (w_1, \ldots, w_5) \in S^*$	**5147.2**	112.2	**5143.3**	**4939.4**	**5375.0**
$(x_1, \ldots, x_5) \in S^*,\ (w_1, \ldots, w_5) \in U^*$	5143.4	112.9	5139.7	4933.8	5372.3

pair of five-factor sets, one drawn from U^* and one from S^*, we compute annualized tangency-portfolio Sharpe ratios both in-sample and out-of-sample, using the second half of the sample as the evaluation window. We also examine the effect of augmenting each block with `FF6`. Table 9 shows that survivor-based specifications dominate on every Sharpe-ratio margin. The average out-of-sample Sharpe ratio of five factors drawn from U^* is 0.86, compared with 0.24 for five factors drawn from S^*. After augmenting each block with `FF6`, the gap widens: the corresponding means are 1.18 and 0.69. Even the full augmentation using all 27 factors in U^* together with `FF6` continues to outperform the matched spanned-set benchmark out of sample. This pattern is consistent with the screening logic: conditional on `FF6`, the anomalies in U^* are precisely those least easily absorbed by the benchmark and thus the most promising candidates for augmenting the pricing kernel.

Table 9 Average Sharpe Ratios of Tangency Portfolios Across Different Factor Configurations.

Note: This table reports the average tangency-portfolio Sharpe ratio over the full sample and the out-of-sample period for various factor configurations. Panel A compares five-factor sets drawn from the survivor pool U^* with equally sized sets drawn from the spanned pool S^*. Panel B augments each block with FF6. Panel C uses all 27 factors in U^* or an equally sized draw from S^*, each augmented with FF6. For S^*, we randomly draw 50,000 combinations of the relevant cardinality from the 126 spanned anomalies. The out-of-sample evaluation uses the second half of the full sample.

f	# Factors	# Combinations	Sharpe Ratio	
			Full sample	OOS
Panel A. Anomalies only				
U^*	5	80,730	**0.98**	**0.86**
S^*	5	50,000	0.60	0.24
Panel B. Anomalies plus benchmark factors				
$(U^*, \texttt{FF6})$	5+6	80,730	**1.71**	**1.18**
$(S^*, \texttt{FF6})$	5+6	50,000	1.38	0.69
Panel C. Full anomaly set plus benchmark factors				
$(U^*, \texttt{FF6})$	27+6	1	**2.26**	**1.01**
$(S^*, \texttt{FF6})$	27+6	50,000	1.93	0.48

We then ask whether the same ranking appears in standard cross-sectional pricing diagnostics. Since Table 9 shows particularly strong performance for $(U^*, \texttt{FF6})$ with five survivors, and because five-anomaly augmentation matches the terminal stage of the selection procedure, we focus on that configuration and compare it with matched five-anomaly draws from S^*. Table 10 reports average explanatory power across three test-asset universes. Across the 49 Industry and 120 Bisort portfolios, $(U^*, \texttt{FF6})$ delivers lower RMS alpha, lower RMSE, lower mean absolute alpha, lower GRS statistics, and higher total R^2 than $(S^*, \texttt{FF6})$. For the 100 P-Tree portfolios the differences are small, but the spanned set does not improve on the survivor set on any metric. Once the benchmark is held fixed, survivor-based augmentations retain stronger pricing ability than equally sized augmentations drawn from S^*.

Table 10 Average Explanatory Power of Different Factor Augmentations Across Test Assets.

Note: This table reports average out-of-sample asset-pricing performance for all factor sets formed by augmenting FF6 with every possible combination of five survivors from U^*, and for 50,000 sets formed by augmenting FF6 with random combinations of five anomalies drawn from S^*. The out-of-sample window is the second half of the full sample. Performance is summarized by root mean square alpha (RMS α), root mean squared error (RMSE $= \sqrt{\frac{1}{NT}\sum_i \sum_t \hat{\varepsilon}_{it}^2}$), mean absolute alpha ($|\alpha| = \frac{1}{N}\sum_i |\hat{\alpha}_i|$), the GRS statistic, and total R^2 ($1 - \frac{\sum_i \sum_t \hat{\varepsilon}_{it}^2}{\sum_i \sum_t r_{it}^2}$). All entries are averages across the relevant factor combinations.

f	# Factors	# Combinations	Metrics				
			RMS α	RMSE	$\lvert\alpha\rvert$	GRS stat.	Total R^2
Panel A. 49 Industry portfolios							
$(U^*, \texttt{FF6})$	11	80,730	**0.0041**	**0.0474**	**0.0032**	**2.10**	**53.2**
$(S^*, \texttt{FF6})$	11	50,000	0.0043	0.0480	0.0033	2.33	52.1
Panel B. 100 P-Tree portfolios							
$(U^*, \texttt{FF6})$	11	80,730	0.0038	0.0387	0.0028	**2.93**	65.8
$(S^*, \texttt{FF6})$	11	50,000	0.0038	0.0387	0.0028	3.07	65.8
Panel C. 120 Bisort portfolios							
$(U^*, \texttt{FF6})$	11	80,730	0.0019	**0.0149**	0.0015	**13.46**	**95.0**
$(S^*, \texttt{FF6})$	11	50,000	0.0019	0.0152	0.0015	14.67	94.8

5.2 Spanned Sets under Different Benchmarks

The results above are benchmark-specific, so the next question is whether the screening outcome is fragile to the benchmark choice. To address this, we rerun the full procedure under an alternative six-factor benchmark, NewB6, that retains only Mkt from FF6. If the spanning classifications were driven mainly by an idiosyncratic benchmark choice, the composition of the spanned set would shift sharply. If, instead, the procedure is identifying robust redundancy, the overlap should remain high.

Specifically, we retain the excess market return (Mkt) but replace the other FF6 factors with the profitability (ROE) and investment (IA) factors from the Hou, Xue, and Zhang (2020) q-factor model, the Pástor and Stambaugh (2003) liquidity factor (LIQ), the Frazzini and Pedersen (2014) betting-against-beta factor (BAB), and an alternative value

factor, HMLD, proposed by Asness and Frazzini (2013). In other words, we continue to use six factors as the benchmark; apart from Mkt, all constituents differ from FF6. We denote this new benchmark by NewB6, which is aligned with Chib, Zeng, and Zhao (2020). [4] The benchmark menu is therefore materially altered even though the dimensionality is unchanged. As before, the key point is that a spanning conclusion does not require the analyst to have chosen the optimal benchmark ex ante.

Figure 5 Counts of Fake Anomalies and Survivors (Benchmark: NewB6).
Note: This figure reports the counts of anomalies classified as $\boldsymbol{w}$ and $\boldsymbol{x}$ by our procedure, using NewB6 as the benchmark. Green bars denote survivors ($\boldsymbol{x}$), and red bars denote fake anomalies ($\boldsymbol{w}$).

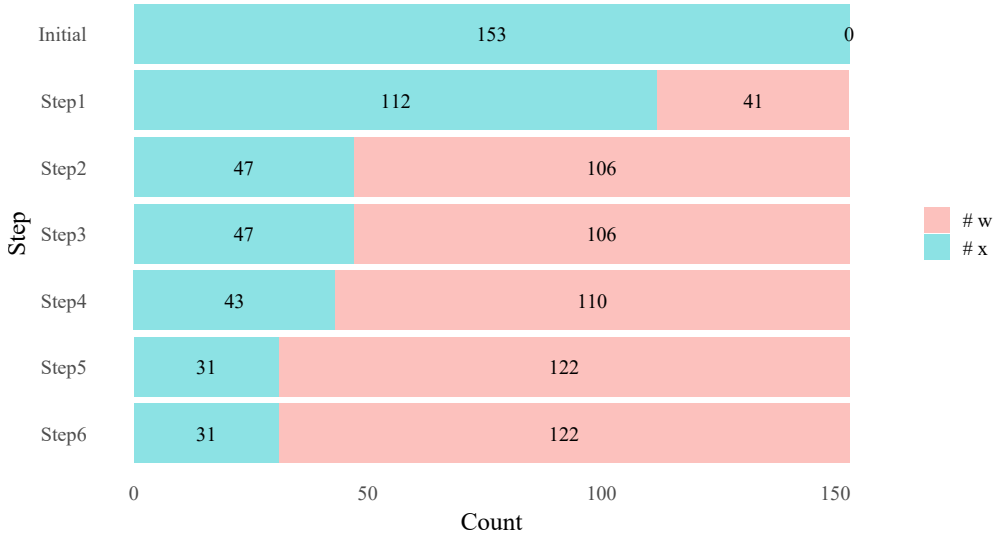
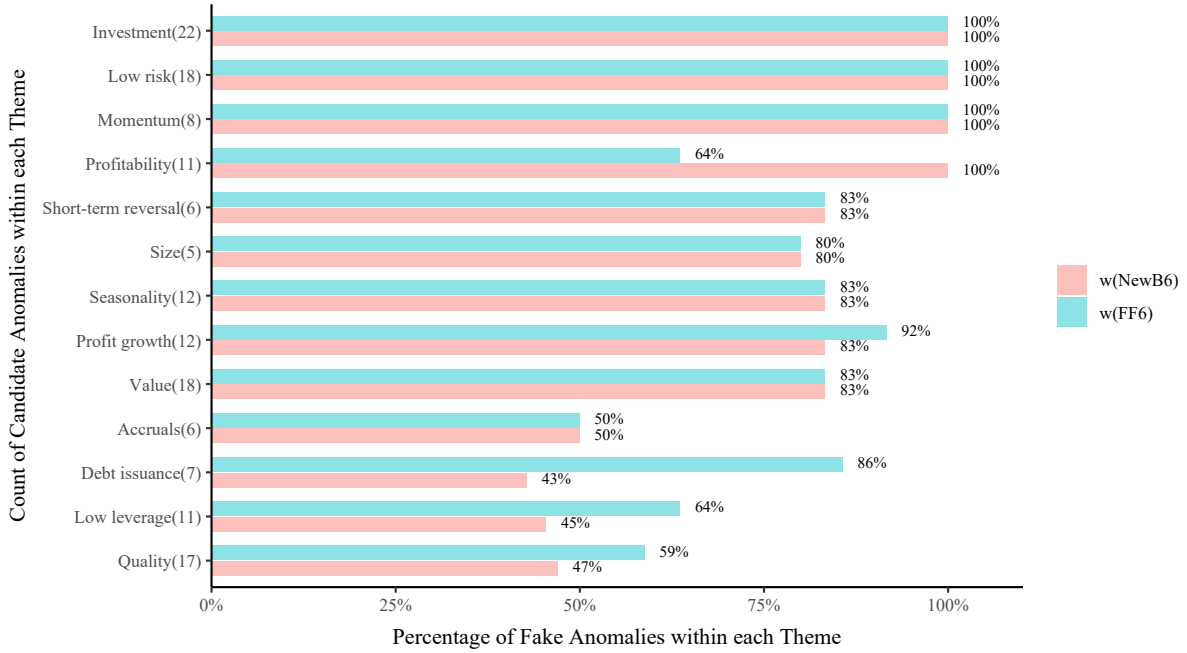


Figure 5 shows that the overall screening path under NewB6 is similar to the path under FF6, although NewB6 classifies slightly fewer anomalies as spanned overall. Figure 6 then shows where the differences come from. Investment, Low risk, Momentum, Seasonality, Short-term reversal, Size, and Accruals are essentially unchanged across the two benchmark choices. The largest shifts are concentrated in Profitability and Debt Issuance. Under NewB6, all profitability anomalies are classified as spanned, which is plausible because ROE directly imports profitability information into the benchmark. By contrast, the share classified as spanned within Debt issuance

[4] Chib et al. (2020) also includes the QMJ factor, but since QMJ is among the 153 candidate anomalies in our sample, we do not include it in the benchmark here.

Figure 6 Percentage of Fake Anomalies within Each Theme.
Note: This figure reports, for each theme category, the percentage of anomalies classified into the spanned set under two benchmarks: FF6 (green) and NewB6 (red). The y-axis lists category names together with the number of candidate anomalies in each category.



falls from 86% to 43%, indicating that the alternative benchmark is a less complete substitute for the financing and investment channels absorbed by CMA, SMB, and MOM.

In fact, although the benchmark matters, the spanned anomalies identified under different benchmarks are highly similar. Table 11 shows that benchmark sensitivity is limited in the aggregate. We summarize overlap from four angles: set similarity (Jaccard index), the precision of FF6 against NewB6 ($\text{Precision}_{\text{FF6}}$), the precision of NewB6 against FF6 ($\text{Precision}_{\text{NewB6}}$), [5] and their harmonic mean (F_1-score). Let $\#\,\boldsymbol{w}_{\text{FF6}}$ and $\#\,\boldsymbol{w}_{\text{NewB6}}$ denote the numbers of fake anomalies under FF6 and NewB6, respectively; and $\#\,\boldsymbol{w}_{\text{Inter}}$ denote the number identified as fake by both benchmarks, i.e., $\boldsymbol{w}_{\text{Inter}} =$

[5] Conventionally, one would label it as the coverage of FF6 by NewB6 ($\text{Recall}_{\text{NewB6}}$). Since the two benchmarks are treated symmetrically here, we instead name such an index as ($\text{Precision}_{\text{NewB6}}$).

Table 11 Degree of Overlap in Spanned Sets under Different Benchmarks.

Note: This table measures the overlap between the spanned sets obtained under the FF6 and NewB6 benchmarks using several complementary metrics. # $\boldsymbol{a}$ denotes the number of candidate anomalies in each theme; # $\boldsymbol{w}_{\texttt{FF6}}$ and # $\boldsymbol{w}_{\texttt{NewB6}}$ denote the numbers of anomalies classified into the spanned set under FF6 and NewB6, respectively; and # $\boldsymbol{w}_{\text{Inter}}$ denotes the number classified as spanned under both benchmarks. The Jaccard index, Precision$_{\texttt{FF6}}$, Precision$_{\texttt{NB6}}$, and the F_1 score summarize the degree of overlap.

Theme	# $\boldsymbol{a}$	# $\boldsymbol{w}_{\texttt{FF6}}$	# $\boldsymbol{w}_{\texttt{NewB6}}$	# $\boldsymbol{w}_{\text{Inter}}$	Jaccard index	Precision$_{\texttt{FF6}}$	Precision$_{\texttt{NB6}}$	F_1-score
Investment	22	22	22	22	1.00	1.00	1.00	1.00
Low risk	18	18	18	18	1.00	1.00	1.00	1.00
Seasonality	12	10	10	10	1.00	1.00	1.00	1.00
Momentum	8	8	8	8	1.00	1.00	1.00	1.00
Short-term reversal	6	5	5	5	1.00	1.00	1.00	1.00
Size	5	4	4	4	1.00	1.00	1.00	1.00
Accruals	6	3	3	3	1.00	1.00	1.00	1.00
Profit growth	12	11	10	10	0.91	0.91	1.00	0.95
Value	18	15	15	14	0.88	0.93	0.93	0.93
Low leverage	11	7	5	5	0.71	0.71	1.00	0.83
Profitability	11	7	11	7	0.64	1.00	0.64	0.78
Quality	17	10	8	7	0.64	0.70	0.88	0.78
Debt issuance	7	6	3	3	0.50	0.50	1.00	0.67
Total	153	126	122	116	0.88	0.92	0.95	0.94

$\boldsymbol{w}_{\texttt{FF6}} \cap \boldsymbol{w}_{\texttt{NewB6}}$. The metrics are computed as follows:

$$\text{Jaccard } J = \frac{\#\boldsymbol{w}_{\text{Inter}}}{\#\boldsymbol{w}_{\texttt{FF6}} + \#\boldsymbol{w}_{\texttt{NewB6}} - \#\boldsymbol{w}_{\text{Inter}}}, \qquad \text{Precision}_{\texttt{FF6}} = \frac{\#\boldsymbol{w}_{\text{Inter}}}{\#\boldsymbol{w}_{\texttt{FF6}}},$$

$$\text{Precision}_{\texttt{NewB6}} = \frac{\#\boldsymbol{w}_{\text{Inter}}}{\#\boldsymbol{w}_{\texttt{NewB6}}}, \qquad F_1\text{-score} = \frac{2\#\boldsymbol{w}_{\text{Inter}}}{\#\boldsymbol{w}_{\texttt{FF6}} + \#\boldsymbol{w}_{\texttt{NewB6}}}.$$

Under FF6 and NewB6, the procedure classifies 126 and 122 anomalies as spanned, respectively, and the intersection of the two spanned sets contains 116 anomalies. The Jaccard index of 0.88 and the F_1 score of 0.94 indicate substantial overlap despite the fact that the two benchmarks share only Mkt. The two precision measures are also close, implying that the similarity is not one-sided. This is an important property of our procedure: benchmark choice changes some marginal classifications, but the bulk of the spanning classifications is stable across reasonable benchmark perturbations.

5.3 Union of Spanned Sets: Robust Survivors

To isolate the anomalies that survive both benchmark exercises, we define $\boldsymbol{w}_{\text{Union}} = \boldsymbol{w}_{\texttt{FF6}} \cup \boldsymbol{w}_{\texttt{NewB6}}$, which contains 132 elements. The complement, denoted $\boldsymbol{x}^*$, contains 21 robust survivors. Because the `FF6`-based screen leaves 27 survivors, the overlap set $\boldsymbol{x}^*$ accounts for 78% of them. The economic interpretation we emphasize below is therefore not driven by a knife-edge benchmark choice; it is anchored in the subset of anomalies that survive under both benchmark menus.

Table 12 Survivors Across Benchmarks.

Note: This table reports the 21 anomalies that remain after taking the union of spanned sets under the benchmarks `FF6` and `NewB6`. For each anomaly, we provide its name, theme, concise description, and literature source.

Names of $\boldsymbol{x}$	Theme	Description	Citation
`cowc_gr1a`	Accruals	Change in current operating working capital	Richardson et al. (2005)
`oaccruals_at`	Accruals	Operating accruals	Sloan (1996)
`oaccruals_ni`	Accruals	Percent operating accruals	Hafzalla et al. (2011)
`noa_at`	Debt issuance	Net operating assets	Hirshleifer et al. (2004)
`cash_at`	Low leverage	Cash-to-assets	Palazzo (2012)
`netdebt_me`	Low leverage	Net debt-to-price	Penman et al. (2007)
`rd_sale`	Low leverage	R&D-to-sales	Chan et al. (2001)
`rd5_at`	Low leverage	R&D capital-to-book assets	Li (2011)
`sale_emp_gr1`	Profit growth	Labor force efficiency	Abarbanell and Bushee (1998)
`cop_at`	Quality	Cash-based operating profits-to-book assets	Ball et al. (2016)
`cop_atl1`	Quality	Cash-based operating profits-to-lagged book assets	Ball et al. (2016)
`gp_at`	Quality	Gross profits-to-assets	Novy-Marx (2013)
`qmj`	Quality	Quality minus Junk: Composite	Asness et al. (2019)
`qmj_safety`	Quality	Quality minus Junk: Safety	Asness et al. (2019)
`sale_bev`	Quality	Assets turnover	Soliman (2008)
`seas_6_10an`	Seasonality	Years 6-10 lagged returns, annual	Heston and Sadka (2008)
`seas_11_15an`	Seasonality	Years 11-15 lagged returns, annual	Heston and Sadka (2008)
`rd_me`	Size	R&D-to-market	Chan et al. (2001)
`rmax5_rvol_21d`	Short-term reversal	Highest 5 days of return scaled by volatility	Asness et al. (2020)
`fcf_me`	Value	Free cash flow-to-price	Lakonishok et al. (1994)
`div12m_me`	Value	Dividend yield	Litzenberger and Ramaswamy (1979)

These 21 survivors coalesce into four economically distinct channels: (i) fundamental quality, (ii) management decisions, (iii) valuation level, and (iv) market behavior and patterns. Each channel highlights a source of risk that the benchmark factor model leaves unspanned or only partially absorbs.

The first channel, *fundamental quality*, captures indicators of firm health, operating

efficiency, and financial robustness, including cash-based profitability (cop_at, cop_atl1), gross profitability (gp_at), composite quality (qmj, qmj_safety), asset turnover (sale_bev), labor productivity (sale_emp_gr1), accrual quality (cowc_gr1a, oaccruals_at, oaccruals_ni), and balance-sheet defensiveness (cash_at, netdebt_me). These measures emphasize cash-flow strength and balance-sheet resilience, reflecting firms' cost structure, operating durability, and ability to sustain value creation. They differ from RMW by emphasizing cash rather than book earnings and by more directly capturing the reliability of accounting information. Relative to HML, they better capture the persistence of profitability and the downside protection associated with superior operations and financial prudence. Although the quality premium is often favored when liquidity is tight, its mechanism does not arise solely from trading costs, so LIQ is at most an incomplete proxy. Thus, anomalies in this channel are more likely to remain unspanned under reasonable benchmark choices.

The second channel, *management decisions*, centers on firms' strategic capital allocation and implications for long-term value. It is captured by two groups of measures: net operating assets (noa_at), which reflect physical investment, and R&D intensity metrics (rd_sale, rd5_at, rd_me), which proxy for innovation effort. Unlike CMA and IA, which place greater emphasis on investment levels or physical investment intensity, this bundle focuses more on the quality and efficiency of investment. Because the value of R&D is only imperfectly reflected on the balance sheet, these exposures also do not fully overlap with HML or CMA. Consequently, anomalies in this channel remain meaningfully likely to stay surviving under benchmarks.

The third channel, *valuation level*, gauges price relative to fundamentals and is proxied by two shareholder cash-flow yields: free-cash-flow yield (fcf_me) and dividend yield (div12m_me). These measures quantify, at the prevailing market capitalization, cash that is actually distributed or feasibly distributable to shareholders. Because they are cash-based, they are less sensitive to noncash accounting charges (for example, depreciation) and more cleanly reflect the firm's value-creation capacity.

The fourth channel, *market behavior and patterns*, abstracts from firm fundamentals and focuses on pricing regularities driven by market microstructure and investor flows. It comprises short-horizon reversal (`rmax5_rvol_21d`), which reflects overreaction to extreme news followed by mean reversion, and long-lag seasonality (`seas_6_10an`, `seas_11_15an`), which is plausibly linked to calendar effects and institutional rebalancing cycles. These signals are not the mirror image of `MOM`: momentum captures intermediate-horizon trends, whereas short-horizon reversal reflects contrarian liquidity provision around temporary order-imbalance shocks, and seasonality reflects systematic timing in flows. These exposures are not a mirror image or simple re-expression of `BAB`. Their persistence indicates departures from frictionless, fully efficient markets; the associated risks, arising from trading frictions, behavioral biases, and flow dynamics, are absorbed only imperfectly by standard factor models.

Taken together, the robust survivor set sharpens the economic map of what the two benchmarks fail to absorb: fundamental quality, corporate policies and financing/investment choices, valuation, and market-behavior channels. We interpret these dimensions as disciplined candidates for benchmark augmentation rather than as a definitive list of primitive SDF components.

5.4 Out-of-sample Evaluation

The benchmark-robust survivor set, together with the benchmark factors, defines a natural menu of candidate SDF augmentations. Essentially, the set formed by the final survivors together with `FF6` (or `NewB6`) constitutes a candidate factor menu that may enter the SDF. We ask whether combinations built from $\boldsymbol{x}^*$ improve portfolio efficiency and whether the posterior market prices of risk isolate a smaller core within the survivor set. Formally, we write:

$$\boldsymbol{x}^* = \underbrace{\boldsymbol{\lambda}_{x^*}}_{(6+21)\times 1} + \boldsymbol{\epsilon}_{x^*,t}, \quad \boldsymbol{\epsilon}_{x^*,t} \sim \mathcal{N}_{27}(\mathbf{0}, \boldsymbol{\Omega}_{x^*}) \tag{25}$$

For simplicity, we present only the results combined with FF6. We first examine the OOS performance of tangency portfolios formed from different combinations of the survivor set.[6] Figure 7 summarizes the resulting OOS Sharpe-ratio distributions. As the number of added survivor factors increases, both the mean and the maximum of the OOS Sharpe-ratio distribution rise, while the share of models that underperform FF6 alone declines. Even a single survivor factor lifts the mean OOS Sharpe ratio above that of the benchmark, and the best five-survivor augmentations reach OOS Sharpe ratios around 1.76. Within this post-selection OOS exercise, the survivor set therefore appears to add economically meaningful portfolio value.

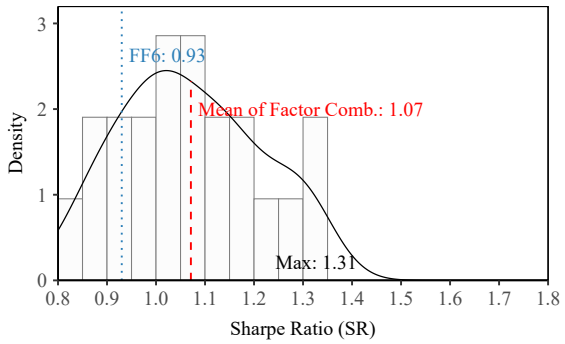
Following Chib et al. (2024), we then recover the marginal posteriors of the market-price vector $\boldsymbol{\Omega}_{x^*}^{-1}\boldsymbol{\lambda}_{x^*}$; Table 13 reports the summaries. On average, absolute market prices are largest for the quality-theme survivors and smallest for the value-theme survivors. Among the benchmark factors, Mkt, RMW, CMA, and MOM have 95% credible intervals that exclude zero, indicating strong evidence that they enter the SDF. Within the survivor set, seas_6_10an, seas_11_15an, rd_me, and rmax5_rvol_21d have 95% credible intervals that exclude zero. Relative to the other survivor factors, these four are the strongest candidates for incremental inclusion in the pricing kernel.

6 Conclusion

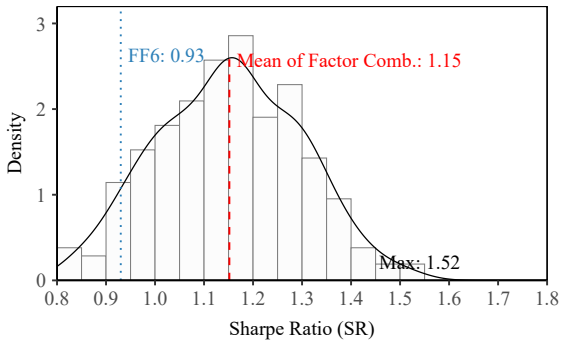
We emphasize a simple but often overlooked point: in an environment of anomaly proliferation and multiple testing, establishing that a signal is not spanned would require knowledge of the true SDF, an object that is essentially unattainable in practice. In contrast, establishing that a signal is spanned only requires exhibiting a factor set that linearly spans it (i.e., zero-intercept pricing). Building on this shift in the burden of proof, we develop a framework centered on Bayesian model partitioning and marginal

[6] Following Chib et al. (2024), for each factor combination, we recursively compute the tangency portfolio for the following month and then evaluate the realized return of the portfolio weights estimated in the previous month. We continue to use the first 25% of the sample to obtain prior information, and we take the final 120 months (January 2015 through December 2024) as the evaluation window.

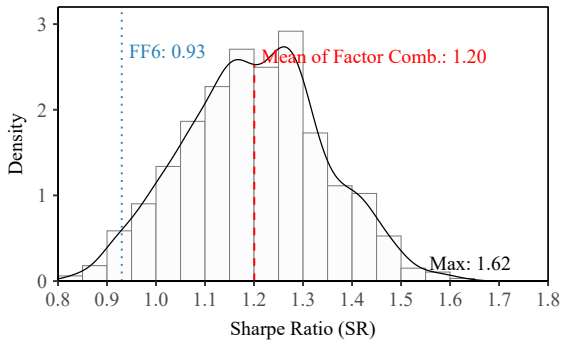
Figure 7 Distributions of Annualized Out-of-Sample Sharpe Ratios.
Note: This figure displays the distributions of the annualized OOS Sharpe ratios for the factor combinations consisting of FF6 and last-stage anomalies based on 10,000 MCMC draws. We consider model classes formed by combining different numbers of $\boldsymbol{x}$ with the benchmark. Specifically, $(1\boldsymbol{x}, \texttt{FF6})$ denotes the 21 models obtained by augmenting FF6 with each single $\boldsymbol{x}$ in turn; for each model, we compute the corresponding OOS Sharpe ratio and then construct the reported statistic. The specifications $(2\boldsymbol{x}, \texttt{FF6})$ and higher orders are defined analogously. The red dashed line denotes the average OOS SR, and the blue dotted line shows the OOS SR obtained when using FF6 alone as the factor model.



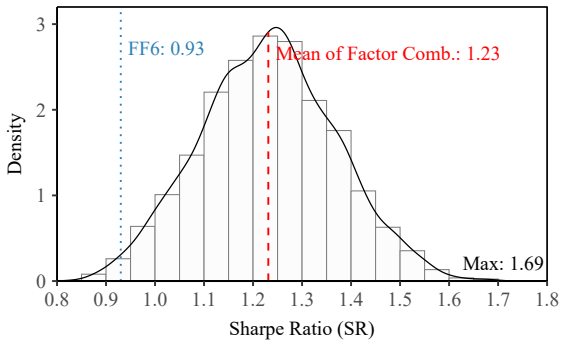
(a) $(1\boldsymbol{x}, \texttt{FF6})$



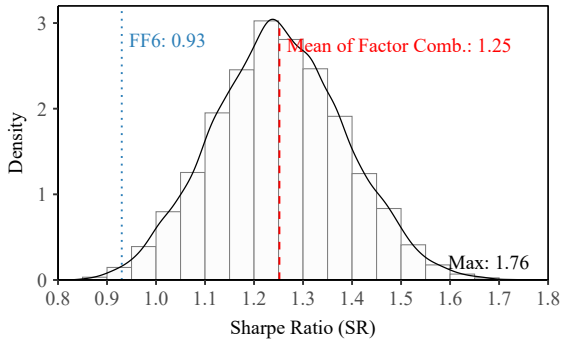
(b) $(2\boldsymbol{x}, \texttt{FF6})$



(c) $(3\boldsymbol{x}, \texttt{FF6})$



(d) $(4\boldsymbol{x}, \texttt{FF6})$



(e) $(5\boldsymbol{x}, \texttt{FF6})$

Table 13 Posterior Statistics of the Market Prices of Factor Risks of FF6 and Survivor Anomalies.

Note: This table reports the posterior statistics (mean, standard deviation, median, and 2.5% and 97.5% quantiles) for the market prices of factor risks, $\boldsymbol{\Omega}_{x^*}^{-1}\boldsymbol{\lambda}_{x^*}$.

Name	Theme	Post. mean	Post. std.	Post. median	2.5% qtile	97.5% qtile
Panel A. Benchmark						
Mkt		6.5	2.1	6.5	**2.4**	**10.7**
SMB		4.9	3.1	4.8	-1.3	11.1
HML		-0.1	4.1	-0.1	-8.2	8.0
RMW		18.4	5.5	18.3	**7.8**	**29.3**
CMA		16.5	5.1	16.5	**6.8**	**26.6**
MOM		4.8	2.0	4.8	**0.9**	**8.7**
Panel B. Survivor Anomalies						
cowc_gr1a	Accruals	1.5	5.0	1.5	-8.3	11.4
oaccruals_at	Accruals	-6.5	7.0	-6.5	-20.2	7.1
oaccruals_ni	Accruals	8.7	6.7	8.6	-4.3	21.8
noa_at	Debt issuance	3.6	5.2	3.6	-6.5	13.8
cash_at	Low leverage	5.8	6.3	5.8	-6.6	18.3
netdebt_me	Low leverage	5.9	6.2	5.9	-6.2	17.9
rd_sale	Low leverage	5.9	5.3	5.9	-4.5	16.3
rd5_at	Low leverage	-10.8	5.5	-10.7	-21.5	-0.2
sale_emp_gr1	Profit growth	-5.4	3.9	-5.4	-13.2	2.2
cop_at	Quality	-0.3	14.2	-0.3	-28.2	27.4
cop_atl1	Quality	18.0	14.4	18.0	-10.1	46.3
gp_at	Quality	-23.7	8.2	-23.7	-40.0	-7.7
qmj	Quality	8.6	6.2	8.6	-3.5	21.0
qmj_safety	Quality	-1.7	5.2	-1.7	-12.0	8.6
sale_bev	Quality	11.0	6.8	11.0	-2.2	24.3
seas_6_10an	Seasonality	8.6	2.6	8.6	**3.5**	**13.8**
seas_11_15an	Seasonality	7.2	3.1	7.1	**1.1**	**13.4**
rd_me	Size	8.7	3.3	8.7	**2.3**	**15.3**
rmax5_rvol_21d	Short-term reversal	6.7	2.4	6.7	**2.0**	**11.6**
fcf_me	Value	4.8	4.0	4.8	-2.9	12.7
div12m_me	Value	-1.3	5.4	-1.3	-11.8	9.3

likelihood, coupled with EFDR control and a stepwise scan. We place benchmark factors and candidate anomalies in a common model space, enumerate all partitions into "unspanned/spanned," and compute for each anomaly its posterior probability of being spanned. We then apply an EFDR threshold and iteratively filter out anomalies that are statistically spanned. In this way, the question moves from "*Is this anomaly an anomaly*" to "*Can this anomaly be spanned?*"

Methodologically, our novel approach addresses two challenges simultaneously. First, model uncertainty. Rather than a binary comparison of the FF benchmark with and without intercept, we integrate over all spanned/unspanned partitions and summarize the evidence via posterior model probabilities. Second, multiplicity. We control EFDR and select the set of spanned anomalies. Crucially, the workflow is deliberately asymmetric toward establishing span, conservative with respect to declaring true anomalies. As a result, failure to be spanned by the current benchmark is not mistaken for evidence of anomaly, and conclusions about the replication crisis do not hinge on a fragile optimal-benchmark assumption.

Using the 153 U.S. factors (Jensen et al., 2023) from 1985 to 2024, our framework delivers clear evidence. With `FF6` as the benchmark, the stepwise scan proceeds as follows: steps 1, 2, 4, and 5 filter 21, 87, 5, and 13 fake anomalies, respectively (step 3 selects none), stabilizing at 126 spanned and 27 unspanned anomalies. These findings are not mere statistical artifacts: under basic frequentist and Bayesian tests, `FF6` plus 1 to 5 of the $\boldsymbol{x}$ explain all 126 fake anomalies. On the investment side, portfolios consisting of `FF6` plus 5 of the last-stage anomalies deliver in-sample and out-of-sample tangency-portfolio Sharpe ratios and cross-sectional fit for industry, bi-sort, and P-Tree test assets that systematically exceed those of fake anomaly sets of the same size.

We find that the cross-benchmark (`FF6` and `NewB6`) union of spanned anomalies equals 132, leaving a residual of 21 anomalies in the survivor set. The intersection of the two benchmark results closely matches the `FF6`-based set, indicating that the economic content we isolate largely corresponds to what `FF6` fails to span and that the finding is robust to reasonable benchmark re-specification. Furthermore, we find eight factors, four from `FF6` and four from the 21 anomalies, whose credible intervals of the posterior distributions of the market prices of risk are strictly positive (excluding 0), indicating that these eight factors are the strongest candidates to enter the true SDF.

Moreover, the 21 survivors cohere along four economically linked yet only partially overlapping channels. The stability and economic coherence of these 21 survivors

indicate that SDF exposures are non-redundant. These deliver state-contingent insurance in high-marginal-utility states and generate verifiable incremental cross-sectional pricing power. As such, they constitute candidates for augmenting the SDF beyond the standard benchmark menu.

In summary, the paper makes three contributions to the anomaly-evaluation literature. First, we transform the untestable claim of anomaly existence into a testable claim about spanning, and supply reproducible decision rules based on posterior model probabilities and EFDR. Second, we introduce an implementable stepwise scan that enables systematic screening in a high-dimensional anomaly zoo. Third, we provide broad-based and robust evidence that most reported signals are spanned in a benchmark-plus-a-few-anomalies setting, while the small set of stable survivors exhibits a coherent economic structure and interpretable risk-friction channels.

References

Abarbanell, J. S. and B. J. Bushee (1998). Abnormal returns to a fundamental analysis strategy. *The Accounting Review*, 19–45.

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

Asness, C., A. Frazzini, N. J. Gormsen, and L. H. Pedersen (2020). Betting against correlation: Testing theories of the low-risk effect. *Journal of Financial Economics 135*(3), 629–652.

Asness, C. S., A. Frazzini, and L. H. Pedersen (2019). Quality minus junk. *Review of Accounting Studies 24*(1), 34–112.

Ball, R., J. Gerakos, J. T. Linnainmaa, and V. Nikolaev (2016). Accruals, cash flows, and operating profitability in the cross section of stock returns. *Journal of Financial Economics 121*(1), 28–45.

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

Chan, L. K., J. Lakonishok, and T. Sougiannis (2001). The stock market valuation of research and development expenditures. *Journal of Finance 56*(6), 2431–2456.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association 90*(432), 1313–1321.

Chib, S. and X. Zeng (2020). Which factors are risk factors in asset pricing? A model scan framework. *Journal of Business & Economic Statistics 38*, 771–783.

Chib, S., X. Zeng, and L. Zhao (2020). On comparing asset pricing models. *Journal of Finance 75*(1), 551–577.

Chib, S., L. Zhao, and G. Zhou (2024). Winners from winners: A tale of risk factors. *Management Science 70*(1), 396–414.

Cochrane, J. H. (2011). Presidential address: Discount rates. *Journal of Finance 66*(4), 1047–1108.

Cong, L. W., G. Feng, J. He, and X. He (2025). Growing the efficient frontier on panel trees. *Journal of Financial Economics 167*, 104024.

Fama, E. F. and K. R. French (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics 33*(1), 3–56.

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

Frazzini, A. and L. H. Pedersen (2014). Betting against beta. *Journal of Financial Economics 111*(1), 1–25.

Hafzalla, N., R. Lundholm, and E. Matthew Van Winkle (2011). Percent accruals. *The Accounting Review 86*(1), 209–236.

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

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

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

Heston, S. L. and R. Sadka (2008). Seasonality in the cross-section of stock returns. *Journal of Financial Economics 87*(2), 418–445.

Hirshleifer, D., K. Hou, S. H. Teoh, and Y. Zhang (2004). Do investors overvalue firms with bloated balance sheets? *Journal of Accounting and Economics 38*, 297–331.

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

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

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

Lakonishok, J., A. Shleifer, and R. W. Vishny (1994). Contrarian investment, extrapolation, and risk. *Journal of Finance 49*(5), 1541–1578.

Lewellen, J., S. Nagel, and J. Shanken (2010). A skeptical appraisal of asset pricing tests. *Journal of Financial Economics 96*(2), 175–194.

Li, F. (2011). Earnings quality based on corporate investment decisions. *Journal of Accounting Research 49*(3), 721–752.

Lintner, J. (1965). The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics 47*, 13–37.

Litzenberger, R. H. and K. Ramaswamy (1979). The effect of personal taxes and dividends on capital asset prices: Theory and empirical evidence. *Journal of Financial Economics 7*(2), 163–195.

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

Müller, P., G. Parmigiani, and K. Rice (2007). FDR and Bayesian multiple comparisons rules. In *Bayesian Statistics 8*, pp. 359–380. Oxford University Press.

Newton, M. A., A. Noueiry, D. Sarkar, and P. Ahlquist (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics 5*(2), 155–176.

Novy-Marx, R. (2013). The other side of value: The gross profitability premium. *Journal of Financial Economics 108*(1), 1–28.

Palazzo, B. (2012). Cash holdings, risk, and expected returns. *Journal of Financial Economics 104*(1), 162–185.

Pástor, L. and R. F. Stambaugh (2003). Liquidity risk and expected stock returns. *Journal of Political Economy 111*(3), 642–685.

Penman, S. H., S. A. Richardson, and I. Tuna (2007). The book-to-price effect in stock returns: Accounting for leverage. *Journal of Accounting Research 45*(2), 427–467.

Richardson, S. A., R. G. Sloan, M. T. Soliman, and I. Tuna (2005). Accrual reliability, earnings persistence and stock prices. *Journal of Accounting and Economics 39*(3), 437–485.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance 19*(3), 425–442.

Sloan, R. G. (1996). Do stock prices fully reflect information in accruals and cash flows about future earnings? *The Accounting Review*, 289–315.

Soliman, M. T. (2008). The use of DuPont analysis by market participants. *The Accounting Review 83*(3), 823–853.

Appendix

Table A.1 Candidate Splits for Factor Decomposition (Benchmark: FF6; one at a time). *Note:* gray (shaded) rows are models where the candidate anomaly a_i is *spanned* (i.e., $a_i \in \boldsymbol{w}$); unshaded rows are models where it is *unspanned* (i.e., $a_i \in \boldsymbol{x}$).

Model	$\boldsymbol{x}$	$\boldsymbol{w}$
$\mathbb{M}_{i,1}$	Mkt	SMB, HML, RMW, CMA, MOM, a_i
$\mathbb{M}_{i,2}$	SMB	Mkt, HML, RMW, CMA, MOM, a_i
$\mathbb{M}_{i,3}$	HML	Mkt, SMB, RMW, CMA, MOM, a_i
$\mathbb{M}_{i,4}$	RMW	Mkt, SMB, HML, CMA, MOM, a_i
$\mathbb{M}_{i,5}$	CMA	Mkt, SMB, HML, RMW, MOM, a_i
$\mathbb{M}_{i,6}$	MOM	Mkt, SMB, HML, RMW, CMA, a_i
$\mathbb{M}_{i,7}$	a_i	Mkt, SMB, HML, RMW, CMA, MOM
$\mathbb{M}_{i,8}$	Mkt, SMB	HML, RMW, CMA, MOM, a_i
$\mathbb{M}_{i,9}$	Mkt, HML	SMB, RMW, CMA, MOM, a_i
$\mathbb{M}_{i,10}$	Mkt, RMW	SMB, HML, CMA, MOM, a_i
$\mathbb{M}_{i,11}$	Mkt, CMA	SMB, HML, RMW, MOM, a_i
$\mathbb{M}_{i,12}$	Mkt, MOM	SMB, HML, RMW, CMA, a_i
$\mathbb{M}_{i,13}$	Mkt, a_i	SMB, HML, RMW, CMA, MOM
$\mathbb{M}_{i,14}$	SMB, HML	Mkt, RMW, CMA, MOM, a_i
$\mathbb{M}_{i,15}$	SMB, RMW	Mkt, HML, CMA, MOM, a_i
$\mathbb{M}_{i,16}$	SMB, CMA	Mkt, HML, RMW, MOM, a_i
$\mathbb{M}_{i,17}$	SMB, MOM	Mkt, HML, RMW, CMA, a_i
$\mathbb{M}_{i,18}$	SMB, a_i	Mkt, HML, RMW, CMA, MOM
$\mathbb{M}_{i,19}$	HML, RMW	Mkt, SMB, CMA, MOM, a_i
$\mathbb{M}_{i,20}$	HML, CMA	Mkt, SMB, RMW, MOM, a_i
$\mathbb{M}_{i,21}$	HML, MOM	Mkt, SMB, RMW, CMA, a_i
$\mathbb{M}_{i,22}$	HML, a_i	Mkt, SMB, RMW, CMA, MOM
$\mathbb{M}_{i,23}$	RMW, CMA	Mkt, SMB, HML, MOM, a_i
$\mathbb{M}_{i,24}$	RMW, MOM	Mkt, SMB, HML, CMA, a_i
$\mathbb{M}_{i,25}$	RMW, a_i	Mkt, SMB, HML, CMA, MOM
$\mathbb{M}_{i,26}$	CMA, MOM	Mkt, SMB, HML, RMW, a_i
$\mathbb{M}_{i,27}$	CMA, a_i	Mkt, SMB, HML, RMW, MOM
$\mathbb{M}_{i,28}$	MOM, a_i	Mkt, SMB, HML, RMW, CMA
$\mathbb{M}_{i,29}$	Mkt, SMB, HML	RMW, CMA, MOM, a_i
$\mathbb{M}_{i,30}$	Mkt, SMB, RMW	HML, CMA, MOM, a_i
$\mathbb{M}_{i,31}$	Mkt, SMB, CMA	HML, RMW, MOM, a_i
$\mathbb{M}_{i,32}$	Mkt, SMB, MOM	HML, RMW, CMA, a_i
$\mathbb{M}_{i,33}$	Mkt, SMB, a_i	HML, RMW, CMA, MOM
$\mathbb{M}_{i,34}$	Mkt, HML, RMW	SMB, CMA, MOM, a_i
$\mathbb{M}_{i,35}$	Mkt, HML, CMA	SMB, RMW, MOM, a_i
$\mathbb{M}_{i,36}$	Mkt, HML, MOM	SMB, RMW, CMA, a_i
$\mathbb{M}_{i,37}$	Mkt, HML, a_i	SMB, RMW, CMA, MOM
$\mathbb{M}_{i,38}$	Mkt, RMW, CMA	SMB, HML, MOM, a_i

continued on next page

Table A.1 (continued)

Model	$\boldsymbol{x}$	$\boldsymbol{w}$
$\mathbb{M}_{i,39}$	Mkt, RMW, MOM	SMB, HML, CMA, a_i
$\mathbb{M}_{i,40}$	Mkt, RMW, a_i	SMB, HML, CMA, MOM
$\mathbb{M}_{i,41}$	Mkt, CMA, MOM	SMB, HML, RMW, a_i
$\mathbb{M}_{i,42}$	Mkt, CMA, a_i	SMB, HML, RMW, MOM
$\mathbb{M}_{i,43}$	Mkt, MOM, a_i	SMB, HML, RMW, CMA
$\mathbb{M}_{i,44}$	SMB, HML, RMW	Mkt, CMA, MOM, a_i
$\mathbb{M}_{i,45}$	SMB, HML, CMA	Mkt, RMW, MOM, a_i
$\mathbb{M}_{i,46}$	SMB, HML, MOM	Mkt, RMW, CMA, a_i
$\mathbb{M}_{i,47}$	SMB, HML, a_i	Mkt, RMW, CMA, MOM
$\mathbb{M}_{i,48}$	SMB, RMW, CMA	Mkt, HML, MOM, a_i
$\mathbb{M}_{i,49}$	SMB, RMW, MOM	Mkt, HML, CMA, a_i
$\mathbb{M}_{i,50}$	SMB, RMW, a_i	Mkt, HML, CMA, MOM
$\mathbb{M}_{i,51}$	SMB, CMA, MOM	Mkt, HML, RMW, a_i
$\mathbb{M}_{i,52}$	SMB, CMA, a_i	Mkt, HML, RMW, MOM
$\mathbb{M}_{i,53}$	SMB, MOM, a_i	Mkt, HML, RMW, CMA
$\mathbb{M}_{i,54}$	HML, RMW, CMA	Mkt, SMB, MOM, a_i
$\mathbb{M}_{i,55}$	HML, RMW, MOM	Mkt, SMB, CMA, a_i
$\mathbb{M}_{i,56}$	HML, RMW, a_i	Mkt, SMB, CMA, MOM
$\mathbb{M}_{i,57}$	HML, CMA, MOM	Mkt, SMB, RMW, a_i
$\mathbb{M}_{i,58}$	HML, CMA, a_i	Mkt, SMB, RMW, MOM
$\mathbb{M}_{i,59}$	HML, MOM, a_i	Mkt, SMB, RMW, CMA
$\mathbb{M}_{i,60}$	RMW, CMA, MOM	Mkt, SMB, HML, a_i
$\mathbb{M}_{i,61}$	RMW, CMA, a_i	Mkt, SMB, HML, MOM
$\mathbb{M}_{i,62}$	RMW, MOM, a_i	Mkt, SMB, HML, CMA
$\mathbb{M}_{i,63}$	CMA, MOM, a_i	Mkt, SMB, HML, RMW
$\mathbb{M}_{i,64}$	Mkt, SMB, HML, RMW	CMA, MOM, a_i
$\mathbb{M}_{i,65}$	Mkt, SMB, HML, CMA	RMW, MOM, a_i
$\mathbb{M}_{i,66}$	Mkt, SMB, HML, MOM	RMW, CMA, a_i
$\mathbb{M}_{i,67}$	Mkt, SMB, HML, a_i	RMW, CMA, MOM
$\mathbb{M}_{i,68}$	Mkt, SMB, RMW, CMA	HML, MOM, a_i
$\mathbb{M}_{i,69}$	Mkt, SMB, RMW, MOM	HML, CMA, a_i
$\mathbb{M}_{i,70}$	Mkt, SMB, RMW, a_i	HML, CMA, MOM
$\mathbb{M}_{i,71}$	Mkt, SMB, CMA, MOM	HML, RMW, a_i
$\mathbb{M}_{i,72}$	Mkt, SMB, CMA, a_i	HML, RMW, MOM
$\mathbb{M}_{i,73}$	Mkt, SMB, MOM, a_i	HML, RMW, CMA
$\mathbb{M}_{i,74}$	Mkt, HML, RMW, CMA	SMB, MOM, a_i
$\mathbb{M}_{i,75}$	Mkt, HML, RMW, MOM	SMB, CMA, a_i
$\mathbb{M}_{i,76}$	Mkt, HML, RMW, a_i	SMB, CMA, MOM
$\mathbb{M}_{i,77}$	Mkt, HML, CMA, MOM	SMB, RMW, a_i
$\mathbb{M}_{i,78}$	Mkt, HML, CMA, a_i	SMB, RMW, MOM
$\mathbb{M}_{i,79}$	Mkt, HML, MOM, a_i	SMB, RMW, CMA
$\mathbb{M}_{i,80}$	Mkt, RMW, CMA, MOM	SMB, HML, a_i
$\mathbb{M}_{i,81}$	Mkt, RMW, CMA, a_i	SMB, HML, MOM
$\mathbb{M}_{i,82}$	Mkt, RMW, MOM, a_i	SMB, HML, CMA

continued on next page

Table A.1 (continued)

Model	$\boldsymbol{x}$	$\boldsymbol{w}$
$\mathbb{M}_{i,83}$	Mkt, CMA, MOM, a_i	SMB, HML, RMW
$\mathbb{M}_{i,84}$	SMB, HML, RMW, CMA	Mkt, MOM, a_i
$\mathbb{M}_{i,85}$	SMB, HML, RMW, MOM	Mkt, CMA, a_i
$\mathbb{M}_{i,86}$	SMB, HML, RMW, a_i	Mkt, CMA, MOM
$\mathbb{M}_{i,87}$	SMB, HML, CMA, MOM	Mkt, RMW, a_i
$\mathbb{M}_{i,88}$	SMB, HML, CMA, a_i	Mkt, RMW, MOM
$\mathbb{M}_{i,89}$	SMB, HML, MOM, a_i	Mkt, RMW, CMA
$\mathbb{M}_{i,90}$	SMB, RMW, CMA, MOM	Mkt, HML, a_i
$\mathbb{M}_{i,91}$	SMB, RMW, CMA, a_i	Mkt, HML, MOM
$\mathbb{M}_{i,92}$	SMB, RMW, MOM, a_i	Mkt, HML, CMA
$\mathbb{M}_{i,93}$	SMB, CMA, MOM, a_i	Mkt, HML, RMW
$\mathbb{M}_{i,94}$	HML, RMW, CMA, MOM	Mkt, SMB, a_i
$\mathbb{M}_{i,95}$	HML, RMW, CMA, a_i	Mkt, SMB, MOM
$\mathbb{M}_{i,96}$	HML, RMW, MOM, a_i	Mkt, SMB, CMA
$\mathbb{M}_{i,97}$	HML, CMA, MOM, a_i	Mkt, SMB, RMW
$\mathbb{M}_{i,98}$	RMW, CMA, MOM, a_i	Mkt, SMB, HML
$\mathbb{M}_{i,99}$	Mkt, SMB, HML, RMW, CMA	MOM, a_i
$\mathbb{M}_{i,100}$	Mkt, SMB, HML, RMW, MOM	CMA, a_i
$\mathbb{M}_{i,101}$	Mkt, SMB, HML, RMW, a_i	CMA, MOM
$\mathbb{M}_{i,102}$	Mkt, SMB, HML, CMA, MOM	RMW, a_i
$\mathbb{M}_{i,103}$	Mkt, SMB, HML, CMA, a_i	RMW, MOM
$\mathbb{M}_{i,104}$	Mkt, SMB, HML, MOM, a_i	RMW, CMA
$\mathbb{M}_{i,105}$	Mkt, SMB, RMW, CMA, MOM	HML, a_i
$\mathbb{M}_{i,106}$	Mkt, SMB, RMW, CMA, a_i	HML, MOM
$\mathbb{M}_{i,107}$	Mkt, SMB, RMW, MOM, a_i	HML, CMA
$\mathbb{M}_{i,108}$	Mkt, SMB, CMA, MOM, a_i	HML, RMW
$\mathbb{M}_{i,109}$	Mkt, HML, RMW, CMA, MOM	SMB, a_i
$\mathbb{M}_{i,110}$	Mkt, HML, RMW, CMA, a_i	SMB, MOM
$\mathbb{M}_{i,111}$	Mkt, HML, RMW, MOM, a_i	SMB, CMA
$\mathbb{M}_{i,112}$	Mkt, HML, CMA, MOM, a_i	SMB, RMW
$\mathbb{M}_{i,113}$	Mkt, RMW, CMA, MOM, a_i	SMB, HML
$\mathbb{M}_{i,114}$	SMB, HML, RMW, CMA, MOM	Mkt, a_i
$\mathbb{M}_{i,115}$	SMB, HML, RMW, CMA, a_i	Mkt, MOM
$\mathbb{M}_{i,116}$	SMB, HML, RMW, MOM, a_i	Mkt, CMA
$\mathbb{M}_{i,117}$	SMB, HML, CMA, MOM, a_i	Mkt, RMW
$\mathbb{M}_{i,118}$	SMB, RMW, CMA, MOM, a_i	Mkt, HML
$\mathbb{M}_{i,119}$	HML, RMW, CMA, MOM, a_i	Mkt, SMB
$\mathbb{M}_{i,120}$	Mkt, SMB, HML, RMW, CMA, MOM	a_i
$\mathbb{M}_{i,121}$	Mkt, SMB, HML, RMW, CMA, a_i	MOM
$\mathbb{M}_{i,122}$	Mkt, SMB, HML, RMW, MOM, a_i	CMA
$\mathbb{M}_{i,123}$	Mkt, SMB, HML, CMA, MOM, a_i	RMW
$\mathbb{M}_{i,124}$	Mkt, SMB, RMW, CMA, MOM, a_i	HML
$\mathbb{M}_{i,125}$	Mkt, HML, RMW, CMA, MOM, a_i	SMB
$\mathbb{M}_{i,126}$	SMB, HML, RMW, CMA, MOM, a_i	Mkt

continued on next page

Table A.1 (continued)

Model	$\boldsymbol{x}$	$\boldsymbol{w}$
$\mathbb{M}_{i,127}$	Mkt, SMB, HML, RMW, CMA, MOM, a_i	$\varnothing$

Table A.2 Candidate Splits for Factor Decomposition (Benchmark: FF6; two at a time). *Note:* gray (shaded) rows are models where the pair (a_{i1}, a_{i2}) *is* jointly spanned ($\{a_{i1}, a_{i2}\} \subseteq \boldsymbol{w}$); unshaded rows are models where they are *jointly unspanned* ($\{a_{i1}, a_{i2}\} \subseteq \boldsymbol{x}$). Mixed configurations are excluded.

Model	$\boldsymbol{x}$	$\boldsymbol{w}$
$\mathbb{M}_{i,1}$	Mkt	SMB, HML, RMW, CMA, MOM, a_{i1}, a_{i2}
$\mathbb{M}_{i,2}$	SMB	Mkt, HML, RMW, CMA, MOM, a_{i1}, a_{i2}
$\mathbb{M}_{i,3}$	HML	Mkt, SMB, RMW, CMA, MOM, a_{i1}, a_{i2}
$\mathbb{M}_{i,4}$	RMW	Mkt, SMB, HML, CMA, MOM, a_{i1}, a_{i2}
$\mathbb{M}_{i,5}$	CMA	Mkt, SMB, HML, RMW, MOM, a_{i1}, a_{i2}
$\mathbb{M}_{i,6}$	MOM	Mkt, SMB, HML, RMW, CMA, a_{i1}, a_{i2}
$\mathbb{M}_{i,7}$	a_{i1}, a_{i2}	Mkt, SMB, HML, RMW, CMA, MOM
$\mathbb{M}_{i,8}$	Mkt, SMB	HML, RMW, CMA, MOM, a_{i1}, a_{i2}
$\mathbb{M}_{i,9}$	Mkt, HML	SMB, RMW, CMA, MOM, a_{i1}, a_{i2}
$\mathbb{M}_{i,10}$	Mkt, RMW	SMB, HML, CMA, MOM, a_{i1}, a_{i2}
$\mathbb{M}_{i,11}$	Mkt, CMA	SMB, HML, RMW, MOM, a_{i1}, a_{i2}
$\mathbb{M}_{i,12}$	Mkt, MOM	SMB, HML, RMW, CMA, a_{i1}, a_{i2}
$\mathbb{M}_{i,13}$	Mkt, a_{i1}, a_{i2}	SMB, HML, RMW, CMA, MOM
$\mathbb{M}_{i,14}$	SMB, HML	Mkt, RMW, CMA, MOM, a_{i1}, a_{i2}
$\mathbb{M}_{i,15}$	SMB, RMW	Mkt, HML, CMA, MOM, a_{i1}, a_{i2}
$\mathbb{M}_{i,16}$	SMB, CMA	Mkt, HML, RMW, MOM, a_{i1}, a_{i2}
$\mathbb{M}_{i,17}$	SMB, MOM	Mkt, HML, RMW, CMA, a_{i1}, a_{i2}
$\mathbb{M}_{i,18}$	SMB, a_{i1}, a_{i2}	Mkt, HML, RMW, CMA, MOM
$\mathbb{M}_{i,19}$	HML, RMW	Mkt, SMB, CMA, MOM, a_{i1}, a_{i2}
$\mathbb{M}_{i,20}$	HML, CMA	Mkt, SMB, RMW, MOM, a_{i1}, a_{i2}
$\mathbb{M}_{i,21}$	HML, MOM	Mkt, SMB, RMW, CMA, a_{i1}, a_{i2}
$\mathbb{M}_{i,22}$	HML, a_{i1}, a_{i2}	Mkt, SMB, RMW, CMA, MOM
$\mathbb{M}_{i,23}$	RMW, CMA	Mkt, SMB, HML, MOM, a_{i1}, a_{i2}
$\mathbb{M}_{i,24}$	RMW, MOM	Mkt, SMB, HML, CMA, a_{i1}, a_{i2}
$\mathbb{M}_{i,25}$	RMW, a_{i1}, a_{i2}	Mkt, SMB, HML, CMA, MOM
$\mathbb{M}_{i,26}$	CMA, MOM	Mkt, SMB, HML, RMW, a_{i1}, a_{i2}
$\mathbb{M}_{i,27}$	CMA, a_{i1}, a_{i2}	Mkt, SMB, HML, RMW, MOM
$\mathbb{M}_{i,28}$	MOM, a_{i1}, a_{i2}	Mkt, SMB, HML, RMW, CMA
$\mathbb{M}_{i,29}$	Mkt, SMB, HML	RMW, CMA, MOM, a_{i1}, a_{i2}
$\mathbb{M}_{i,30}$	Mkt, SMB, RMW	HML, CMA, MOM, a_{i1}, a_{i2}
$\mathbb{M}_{i,31}$	Mkt, SMB, CMA	HML, RMW, MOM, a_{i1}, a_{i2}
$\mathbb{M}_{i,32}$	Mkt, SMB, MOM	HML, RMW, CMA, a_{i1}, a_{i2}
$\mathbb{M}_{i,33}$	Mkt, SMB, a_{i1}, a_{i2}	HML, RMW, CMA, MOM
$\mathbb{M}_{i,34}$	Mkt, HML, RMW	SMB, CMA, MOM, a_{i1}, a_{i2}

continued on next page

Table A.2 (continued)

Model	$\boldsymbol{x}$	$\boldsymbol{w}$
$\mathbb{M}_{i,35}$	Mkt, HML, CMA	SMB, RMW, MOM, a_{i1}, a_{i2}
$\mathbb{M}_{i,36}$	Mkt, HML, MOM	SMB, RMW, CMA, a_{i1}, a_{i2}
$\mathbb{M}_{i,37}$	Mkt, HML, a_{i1}, a_{i2}	SMB, RMW, CMA, MOM
$\mathbb{M}_{i,38}$	Mkt, RMW, CMA	SMB, HML, MOM, a_{i1}, a_{i2}
$\mathbb{M}_{i,39}$	Mkt, RMW, MOM	SMB, HML, CMA, a_{i1}, a_{i2}
$\mathbb{M}_{i,40}$	Mkt, RMW, a_{i1}, a_{i2}	SMB, HML, CMA, MOM
$\mathbb{M}_{i,41}$	Mkt, CMA, MOM	SMB, HML, RMW, a_{i1}, a_{i2}
$\mathbb{M}_{i,42}$	Mkt, CMA, a_{i1}, a_{i2}	SMB, HML, RMW, MOM
$\mathbb{M}_{i,43}$	Mkt, MOM, a_{i1}, a_{i2}	SMB, HML, RMW, CMA
$\mathbb{M}_{i,44}$	SMB, HML, RMW	Mkt, CMA, MOM, a_{i1}, a_{i2}
$\mathbb{M}_{i,45}$	SMB, HML, CMA	Mkt, RMW, MOM, a_{i1}, a_{i2}
$\mathbb{M}_{i,46}$	SMB, HML, MOM	Mkt, RMW, CMA, a_{i1}, a_{i2}
$\mathbb{M}_{i,47}$	SMB, HML, a_{i1}, a_{i2}	Mkt, RMW, CMA, MOM
$\mathbb{M}_{i,48}$	SMB, RMW, CMA	Mkt, HML, MOM, a_{i1}, a_{i2}
$\mathbb{M}_{i,49}$	SMB, RMW, MOM	Mkt, HML, CMA, a_{i1}, a_{i2}
$\mathbb{M}_{i,50}$	SMB, RMW, a_{i1}, a_{i2}	Mkt, HML, CMA, MOM
$\mathbb{M}_{i,51}$	SMB, CMA, MOM	Mkt, HML, RMW, a_{i1}, a_{i2}
$\mathbb{M}_{i,52}$	SMB, CMA, a_{i1}, a_{i2}	Mkt, HML, RMW, MOM
$\mathbb{M}_{i,53}$	SMB, MOM, a_{i1}, a_{i2}	Mkt, HML, RMW, CMA
$\mathbb{M}_{i,54}$	HML, RMW, CMA	Mkt, SMB, MOM, a_{i1}, a_{i2}
$\mathbb{M}_{i,55}$	HML, RMW, MOM	Mkt, SMB, CMA, a_{i1}, a_{i2}
$\mathbb{M}_{i,56}$	HML, RMW, a_{i1}, a_{i2}	Mkt, SMB, CMA, MOM
$\mathbb{M}_{i,57}$	HML, CMA, MOM	Mkt, SMB, RMW, a_{i1}, a_{i2}
$\mathbb{M}_{i,58}$	HML, CMA, a_{i1}, a_{i2}	Mkt, SMB, RMW, MOM
$\mathbb{M}_{i,59}$	HML, MOM, a_{i1}, a_{i2}	Mkt, SMB, RMW, CMA
$\mathbb{M}_{i,60}$	RMW, CMA, MOM	Mkt, SMB, HML, a_{i1}, a_{i2}
$\mathbb{M}_{i,61}$	RMW, CMA, a_{i1}, a_{i2}	Mkt, SMB, HML, MOM
$\mathbb{M}_{i,62}$	RMW, MOM, a_{i1}, a_{i2}	Mkt, SMB, HML, CMA
$\mathbb{M}_{i,63}$	CMA, MOM, a_{i1}, a_{i2}	Mkt, SMB, HML, RMW
$\mathbb{M}_{i,64}$	Mkt, SMB, HML, RMW	CMA, MOM, a_{i1}, a_{i2}
$\mathbb{M}_{i,65}$	Mkt, SMB, HML, CMA	RMW, MOM, a_{i1}, a_{i2}
$\mathbb{M}_{i,66}$	Mkt, SMB, HML, MOM	RMW, CMA, a_{i1}, a_{i2}
$\mathbb{M}_{i,67}$	Mkt, SMB, HML, a_{i1}, a_{i2}	RMW, CMA, MOM
$\mathbb{M}_{i,68}$	Mkt, SMB, RMW, CMA	HML, MOM, a_{i1}, a_{i2}
$\mathbb{M}_{i,69}$	Mkt, SMB, RMW, MOM	HML, CMA, a_{i1}, a_{i2}
$\mathbb{M}_{i,70}$	Mkt, SMB, RMW, a_{i1}, a_{i2}	HML, CMA, MOM
$\mathbb{M}_{i,71}$	Mkt, SMB, CMA, MOM	HML, RMW, a_{i1}, a_{i2}
$\mathbb{M}_{i,72}$	Mkt, SMB, CMA, a_{i1}, a_{i2}	HML, RMW, MOM
$\mathbb{M}_{i,73}$	Mkt, SMB, MOM, a_{i1}, a_{i2}	HML, RMW, CMA
$\mathbb{M}_{i,74}$	Mkt, HML, RMW, CMA	SMB, MOM, a_{i1}, a_{i2}
$\mathbb{M}_{i,75}$	Mkt, HML, RMW, MOM	SMB, CMA, a_{i1}, a_{i2}
$\mathbb{M}_{i,76}$	Mkt, HML, RMW, a_{i1}, a_{i2}	SMB, CMA, MOM
$\mathbb{M}_{i,77}$	Mkt, HML, CMA, MOM	SMB, RMW, a_{i1}, a_{i2}
$\mathbb{M}_{i,78}$	Mkt, HML, CMA, a_{i1}, a_{i2}	SMB, RMW, MOM

continued on next page

Table A.2 (continued)

Model	$\boldsymbol{x}$	$\boldsymbol{w}$
$\mathbb{M}_{i,79}$	Mkt, HML, MOM, a_{i1}, a_{i2}	SMB, RMW, CMA
$\mathbb{M}_{i,80}$	Mkt, RMW, CMA, MOM	SMB, HML, a_{i1}, a_{i2}
$\mathbb{M}_{i,81}$	Mkt, RMW, CMA, a_{i1}, a_{i2}	SMB, HML, MOM
$\mathbb{M}_{i,82}$	Mkt, RMW, MOM, a_{i1}, a_{i2}	SMB, HML, CMA
$\mathbb{M}_{i,83}$	Mkt, CMA, MOM, a_{i1}, a_{i2}	SMB, HML, RMW
$\mathbb{M}_{i,84}$	SMB, HML, RMW, CMA	Mkt, MOM, a_{i1}, a_{i2}
$\mathbb{M}_{i,85}$	SMB, HML, RMW, MOM	Mkt, CMA, a_{i1}, a_{i2}
$\mathbb{M}_{i,86}$	SMB, HML, RMW, a_{i1}, a_{i2}	Mkt, CMA, MOM
$\mathbb{M}_{i,87}$	SMB, HML, CMA, MOM	Mkt, RMW, a_{i1}, a_{i2}
$\mathbb{M}_{i,88}$	SMB, HML, CMA, a_{i1}, a_{i2}	Mkt, RMW, MOM
$\mathbb{M}_{i,89}$	SMB, HML, MOM, a_{i1}, a_{i2}	Mkt, RMW, CMA
$\mathbb{M}_{i,90}$	SMB, RMW, CMA, MOM	Mkt, HML, a_{i1}, a_{i2}
$\mathbb{M}_{i,91}$	SMB, RMW, CMA, a_{i1}, a_{i2}	Mkt, HML, MOM
$\mathbb{M}_{i,92}$	SMB, RMW, MOM, a_{i1}, a_{i2}	Mkt, HML, CMA
$\mathbb{M}_{i,93}$	SMB, CMA, MOM, a_{i1}, a_{i2}	Mkt, HML, RMW
$\mathbb{M}_{i,94}$	HML, RMW, CMA, MOM	Mkt, SMB, a_{i1}, a_{i2}
$\mathbb{M}_{i,95}$	HML, RMW, CMA, a_{i1}, a_{i2}	Mkt, SMB, MOM
$\mathbb{M}_{i,96}$	HML, RMW, MOM, a_{i1}, a_{i2}	Mkt, SMB, CMA
$\mathbb{M}_{i,97}$	HML, CMA, MOM, a_{i1}, a_{i2}	Mkt, SMB, RMW
$\mathbb{M}_{i,98}$	RMW, CMA, MOM, a_{i1}, a_{i2}	Mkt, SMB, HML
$\mathbb{M}_{i,99}$	Mkt, SMB, HML, RMW, CMA	MOM, a_{i1}, a_{i2}
$\mathbb{M}_{i,100}$	Mkt, SMB, HML, RMW, MOM	CMA, a_{i1}, a_{i2}
$\mathbb{M}_{i,101}$	Mkt, SMB, HML, RMW, a_{i1}, a_{i2}	CMA, MOM
$\mathbb{M}_{i,102}$	Mkt, SMB, HML, CMA, MOM	RMW, a_{i1}, a_{i2}
$\mathbb{M}_{i,103}$	Mkt, SMB, HML, CMA, a_{i1}, a_{i2}	RMW, MOM
$\mathbb{M}_{i,104}$	Mkt, SMB, HML, MOM, a_{i1}, a_{i2}	RMW, CMA
$\mathbb{M}_{i,105}$	Mkt, SMB, RMW, CMA, MOM	HML, a_{i1}, a_{i2}
$\mathbb{M}_{i,106}$	Mkt, SMB, RMW, CMA, a_{i1}, a_{i2}	HML, MOM
$\mathbb{M}_{i,107}$	Mkt, SMB, RMW, MOM, a_{i1}, a_{i2}	HML, CMA
$\mathbb{M}_{i,108}$	Mkt, SMB, CMA, MOM, a_{i1}, a_{i2}	HML, RMW
$\mathbb{M}_{i,109}$	Mkt, HML, RMW, CMA, MOM	SMB, a_{i1}, a_{i2}
$\mathbb{M}_{i,110}$	Mkt, HML, RMW, CMA, a_{i1}, a_{i2}	SMB, MOM
$\mathbb{M}_{i,111}$	Mkt, HML, RMW, MOM, a_{i1}, a_{i2}	SMB, CMA
$\mathbb{M}_{i,112}$	Mkt, HML, CMA, MOM, a_{i1}, a_{i2}	SMB, RMW
$\mathbb{M}_{i,113}$	Mkt, RMW, CMA, MOM, a_{i1}, a_{i2}	SMB, HML
$\mathbb{M}_{i,114}$	SMB, HML, RMW, CMA, MOM	Mkt, a_{i1}, a_{i2}
$\mathbb{M}_{i,115}$	SMB, HML, RMW, CMA, a_{i1}, a_{i2}	Mkt, MOM
$\mathbb{M}_{i,116}$	SMB, HML, RMW, MOM, a_{i1}, a_{i2}	Mkt, CMA
$\mathbb{M}_{i,117}$	SMB, HML, CMA, MOM, a_{i1}, a_{i2}	Mkt, RMW
$\mathbb{M}_{i,118}$	SMB, RMW, CMA, MOM, a_{i1}, a_{i2}	Mkt, HML
$\mathbb{M}_{i,119}$	HML, RMW, CMA, MOM, a_{i1}, a_{i2}	Mkt, SMB
$\mathbb{M}_{i,120}$	Mkt, SMB, HML, RMW, CMA, MOM	a_{i1}, a_{i2}
$\mathbb{M}_{i,121}$	Mkt, SMB, HML, RMW, CMA, a_{i1}, a_{i2}	MOM
$\mathbb{M}_{i,122}$	Mkt, SMB, HML, RMW, MOM, a_{i1}, a_{i2}	CMA

continued on next page

Table A.2 (continued)

Model	$\boldsymbol{x}$	$\boldsymbol{w}$
$\mathbb{M}_{i,123}$	Mkt, SMB, HML, CMA, MOM, a_{i1}, a_{i2}	RMW
$\mathbb{M}_{i,124}$	Mkt, SMB, RMW, CMA, MOM, a_{i1}, a_{i2}	HML
$\mathbb{M}_{i,125}$	Mkt, HML, RMW, CMA, MOM, a_{i1}, a_{i2}	SMB
$\mathbb{M}_{i,126}$	SMB, HML, RMW, CMA, MOM, a_{i1}, a_{i2}	Mkt
$\mathbb{M}_{i,127}$	Mkt, SMB, HML, RMW, CMA, MOM, a_{i1}, a_{i2}	$\varnothing$

Table A.3 Equity Characteristics

Note: This table provides descriptions of 20 characteristics used in the empirical study.

No.	Characteristics	Description	Category
1	ABR	Abnormal returns around earnings announcement	Momentum
2	ACC	Operating accruals	Investment
3	ADM	Advertising expense-to-market	Intangibles
4	AGR	Asset growth	Investment
5	BASPREAD	Bid-ask spread (3 months)	Frictions
6	BETA	Beta (3 months)	Frictions
7	BM	Book-to-market equity	Value-versus-growth
8	CFP	Cashflow-to-price	Value-versus-growth
9	EP	Earnings-to-price	Value-versus-growth
10	ME	Market equity	Frictions
11	MOM1M	Previous month return	Momentum
12	MOM12M	Cumulative returns in the past (2-12) months	Momentum
13	NI	Net equity issue	Investment
14	OP	Operating profitability	Profitability
15	RDM	R&D-to-market	Intangibles
16	ROE	Return on equity	Profitability
17	SEAS1A	1-Year Seasonality	Intangibles
18	SP	Sales-to-price	Value-versus-growth
19	SUE	Standardized unexpected quarterly earnings	Momentum
20	SVAR	Return variance (3 months)	Frictions

Figure A.1 Sensitivity of Anomaly Significance to Benchmark Model Selection

Note: This figure displays the *t*-statistics of the intercepts (α) from time-series regressions of each anomaly return on four benchmark models: CAPM, FF3, FF5, and FF6. Each cell corresponds to the *t*-statistic for a given anomaly-benchmark pair. Red shades indicate positive *t*-statistics, blue shades indicate negative *t*-statistics, and darker colors indicate larger absolute magnitudes.

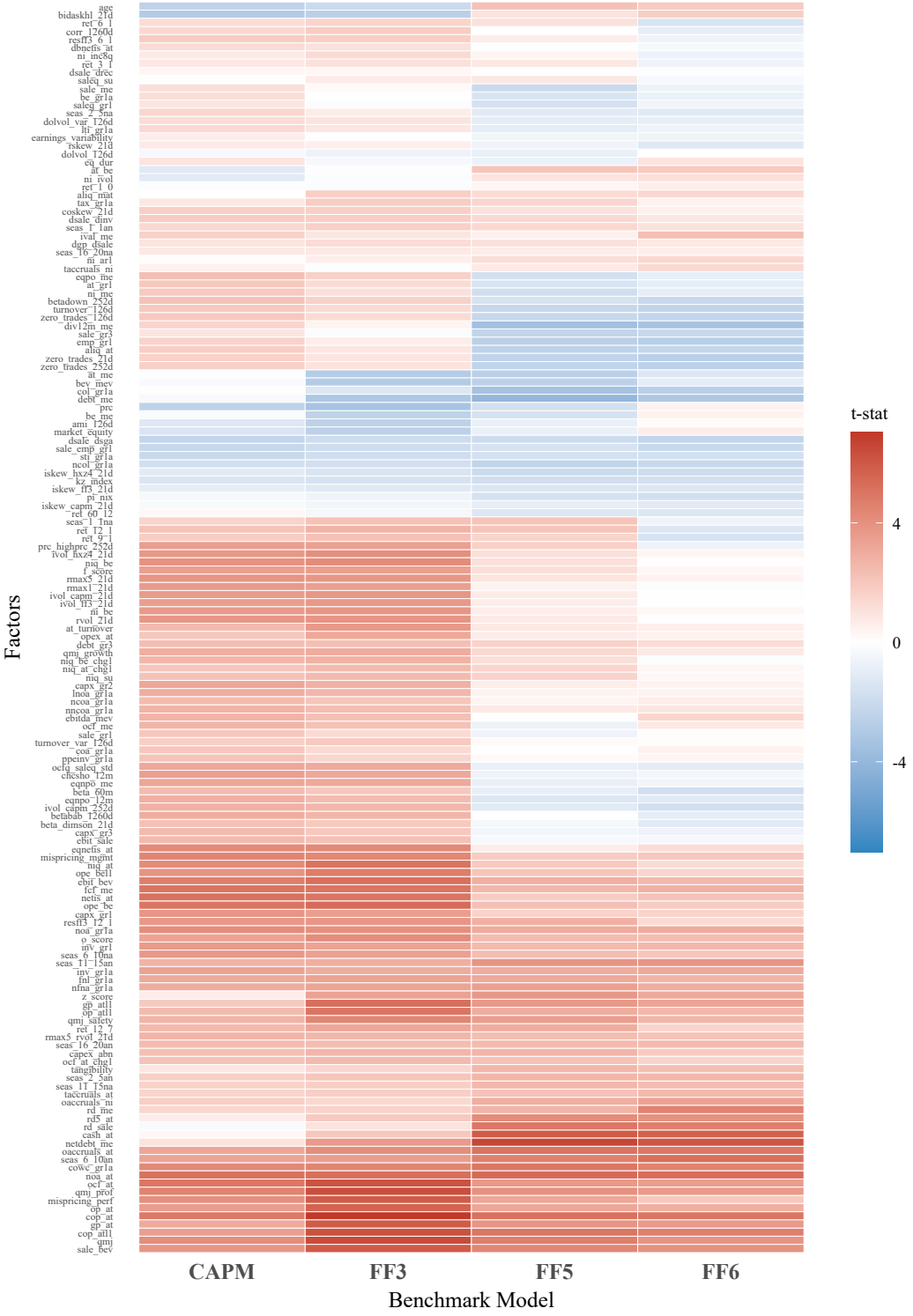


Figure A.2 Counts of Fake Anomalies identified under different q thresholds (in $q = 0.2$ case)
This figure plots, for Steps 1-6, the number of fake anomalies identified when applying EFDR thresholds of $q = 0.10$, $q = 0.15$, and $q = 0.20$, using the EFDR distributions obtained under the baseline configuration with $q = 0.20$.

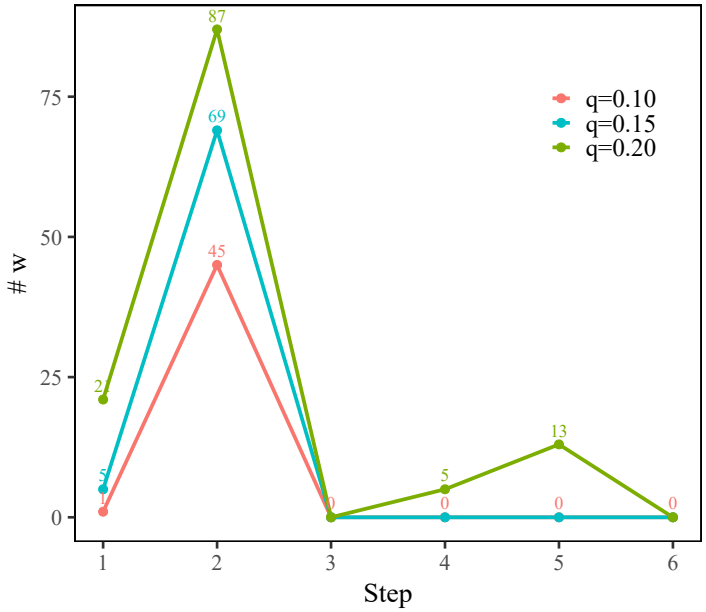


Figure A.3 Counts of Fake Anomalies and Remaining Anomalies. (Benchmark: `FF6`, $q = 0.1$).
Note: This figure reports the counts of anomalies classified as $\boldsymbol{w}$ and $\boldsymbol{x}$ by our procedure, using `FF6` as the benchmark. Green bars denote remaining anomalies ($\boldsymbol{x}$), and red bars denote fake anomalies ($\boldsymbol{w}$). The threshold q is 0.1.

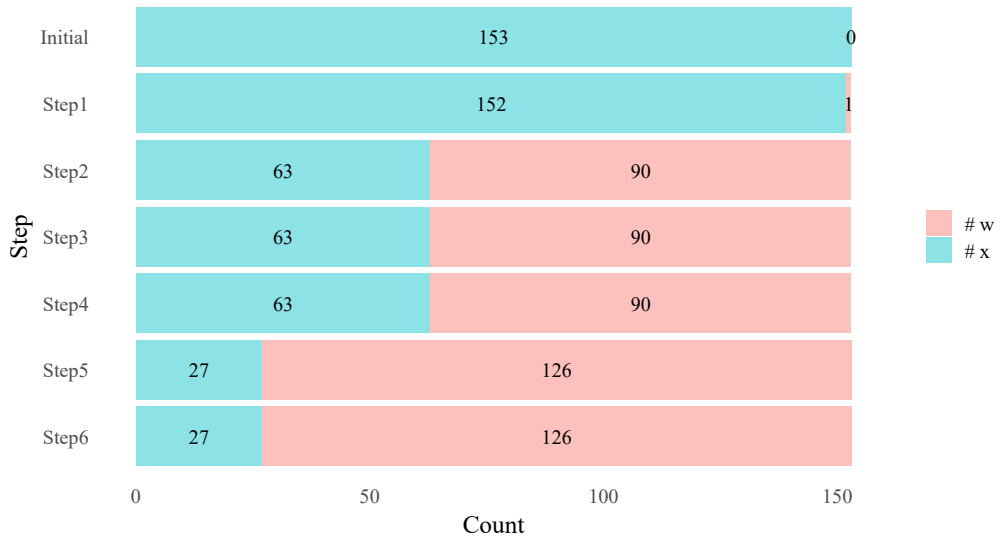


Figure A.4 Bar plot of EFDR and lfdr in the first step.
Note: This figure plots EFDR and lfdr for each anomaly in Step 1 of our procedure. The red bars show EFDR, and the green bars show lfdr.

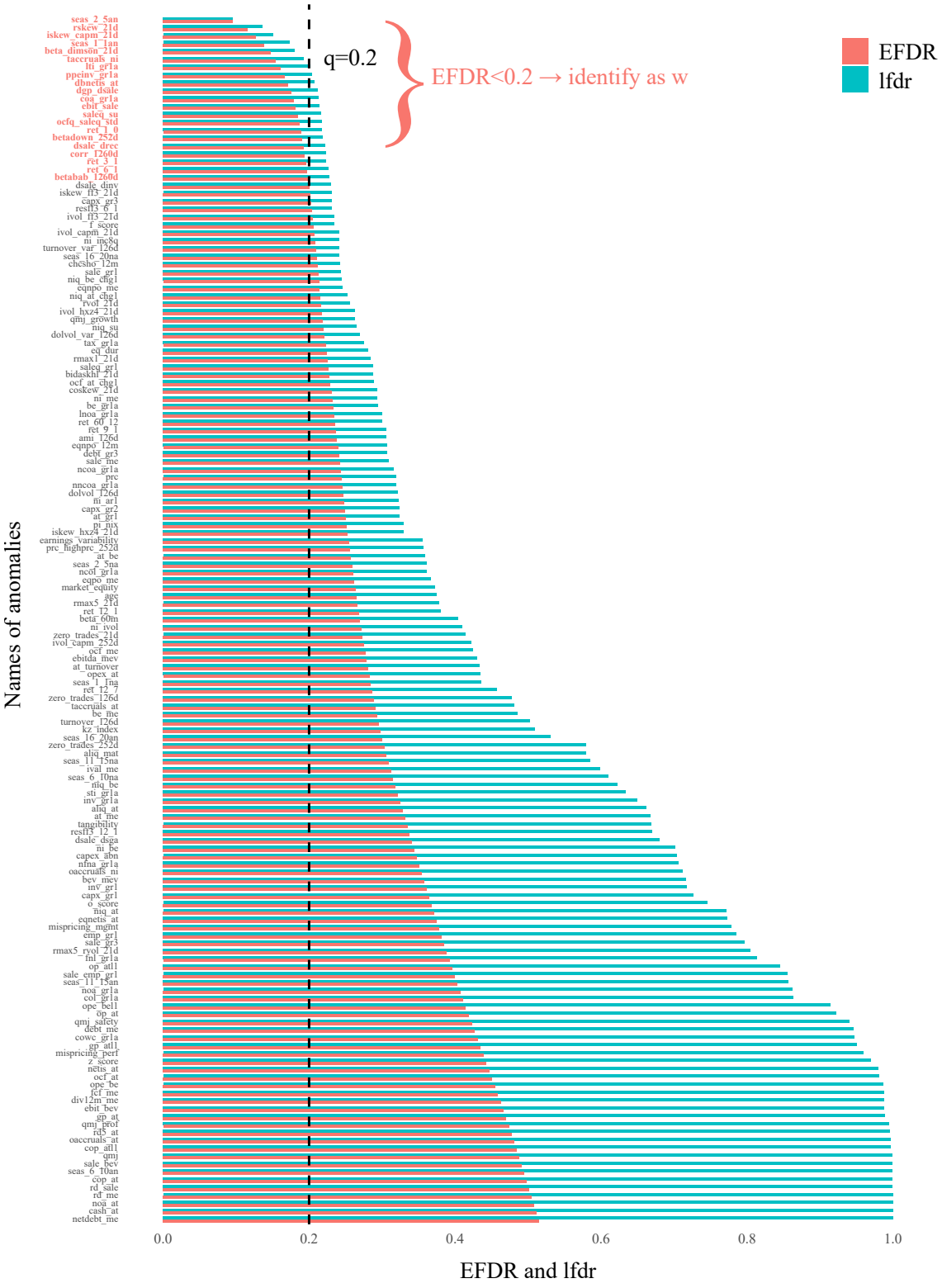


Figure A.5 Eigenvalue Profiles of Spanned Sets.

Note: This figure displays the scree plots for the covariance matrices of anomaly sets classified as spanned (w) at different stages of the stepwise procedure. The horizontal axis represents the rank of the principal component (ordered by eigenvalue magnitude), and the vertical axis reports the corresponding eigenvalue. The lines represent different sets of spanned anomalies: w_1 (Step 1), w_2 (Step 2), w_4 (Step 4), and w_5 (Step 5) for comparison. The number of anomalies (n) in each set is indicated in the legend.