Factor Zoo Revisited: Multiple Testing, Hierarchical Modeling, and Out-of-Sample Evidence

Paulo Roberto Fonteles Guimarães[1], Herbert Kimura[1]

Abstract

The proliferation of return predictors has raised concerns about multiple testing, p-hacking, and out-of-sample performance. While most evidence comes from developed markets, Emerging Markets (EM) provide a demanding out-of-sample environment, where many factors were not originally proposed and face harsher inferential conditions. This paper revisits existence, magnitude, and robustness of asset-pricing discoveries through multiple-testing adjustments and hierarchical modeling.

We examine three questions: whether alphas persist after multiple-testing correction, whether hierarchical theme-based modeling improves inference, and whether these approaches enhance out-of-sample portfolio performance. Using an EM equity panel, we consider frequentist controls and Bayesian hierarchical models, which are embedded in walk-forward backtests.

Results show that only a limited set of themes delivers robust alphas. Hierarchical models reveal economically meaningful theme-level effects while supporting more parsimonious inference. Portfolio-wise, ignoring multiple testing harms performance, while overly conservative frequentist corrections raise risk. FDR-based and hierarchical Bayesian approaches provide a more balanced, economically meaningful framework.

Keywords: factor zoo, asset pricing, multiple testing, partial pooling, emerging markets, hierarchical models, Bayesian statistics, False Discovery Rate, Family-Wise Error Rate

Email addresses: pauloguimaraes871@gmail.com (Paulo Roberto Fonteles Guimarães), herbert.kimura@gmail.com (Herbert Kimura)

1. Introduction

Over the past three decades, the empirical asset pricing literature has documented a vast number of return predictors, commonly referred to as the *factor zoo* (Cochrane, 2011). While many of these factors [1] are individually statistically significant, their joint interpretation raises a fundamental inferential challenge: when hundreds of hypotheses are tested in parallel, standard significance thresholds mechanically generate false discoveries.

A large body of recent research has shown that many documented anomalies exhibit substantial post-publication decay, particularly in developed markets, suggesting that a non-trivial fraction of reported results may reflect data mining, specification search, or crowding effects rather than persistent risk premia (McLean and Pontiff, 2016; Jacobs and Müller, 2020). Several structural features of the literature are appointed as reasons for this problem: strong incentives for publishing new "discoveries" rather than replications, the combination of expanding computational power with relatively limited datasets, and the absence of a clear Popperian-style causal framework for observational empirical research. Collectively, these factors contribute to a phenomenon often described as p-hacking, undermining the credibility of reported factor premia and motivating a credibility crisis in the asset pricing literature (Harvey et al., 2016; De Prado, 2023; Jensen et al., 2023). As a result, controlling for false discoveries has become a central concern in modern asset pricing, with proposed solutions ranging from frequentist error-rate control and Bayesian shrinkage to the application of causal methods.

It is important to notice that, although a widely discussed topic, the severity of the p-hacking problem is an ongoing discussion. Some authors argue that the factor zoo should be viewed under a hierarchical structure as a collection of themes with a higher concentration of academic effort around those with strongest evidence and more plausible theoretical foundations. Given the inherent noise of returns and measurement errors, a partial pooled hierarchical structure should represent a better description of the factor zoo and some frequentist multiple testing adjustments might have the consequence of

[1] A "factor" is often viewed as a common risk source that explain cross-sectional returns through risk premia, rewarding investors for their exposure, while an "anomaly" offers higher average returns regardless of risk exposures. In the present paper, we will not be strict to such definitions, viewing both factors and anomalies more generally as expected return predictors.

significantly increasing Type II error rate and reducing diversification across individual signals (Aghassi et al., 2023; Chen, 2022; Jensen et al., 2023).

In this context, Emerging Markets (EM) offer a particularly relevant and underexplored laboratory for studying the factor zoo. First, most prominent equity factors were originally proposed and validated in developed markets, especially the U.S., making EM a natural out-of-sample environment to assess their external validity. Second, EM are characterized by noisier data, shorter effective samples, higher idiosyncratic volatility, and greater structural heterogeneity, all of which exacerbate the multiple-testing problem and raise the cost of false discoveries. Third, recent evidence suggests that factor crowding and post-publication decay are substantially weaker outside the U.S., implying that EM may have specific dynamics and preserve economically meaningful factor premia even when developed-market evidence has attenuated (Jacobs and Müller, 2020). Together, these features make EM an ideal setting for re-assessing the existence, magnitude, and robustness of factor alphas under disciplined inferential frameworks.

This paper addresses three closely related questions. First, *do factor alphas exist in Emerging Markets, and if so, what is their economically relevant magnitude once multiple testing is properly accounted for?* Second, *does a hierarchical view of the factor zoo, recognizing that individual signals belong to broader thematic groups, improve statistical inference relative to treating factors as independent hypotheses?* Third, *can such hierarchical and multiple-testing-aware frameworks be translated into superior out-of-sample portfolio performance through disciplined factor selection strategies?*

To answer these questions, we develop a framework that considers both frequentist and Bayesian approaches to multiple testing, while also allowing a hierarchical modeling of CAPM alphas. In this context, we evaluate both standard no-pooling anomaly-by-anomaly and partial pooled hierarchical regressions, alongside multiple-testing adjustments that control either the family-wise error rate (FWER) or the false discovery rate (FDR). On the hierarchical side, we estimate Bayesian and frequentist models in which factor alphas are partially pooled within economically motivated themes, allowing information to be shared across related signals while naturally regularizing extreme estimates. For the Bayesian case, we consider both conservative and informative prior specifications, the latter designed to reflect external global factor evidence, especially helpful to stabilize inference in data-scarce environments such as EM.

A central conceptual contribution of the paper is to explicitly link sta-

tistical inference to economic relevance. Rather than evaluating factor significance in isolation, we embed each model's correspondent selection rule (frequentist or Bayesian) into a walk-forward, expanding-window backtesting framework. At each rebalancing date, factors are selected based solely on information available at that time, portfolios are formed using transparent weighting schemes, and performance is evaluated strictly out-of-sample. This design allows us to assess not only whether different inferential frameworks reduce false discoveries, but also whether they lead to economically meaningful improvements in realized returns, risk, and drawdown behavior. Additionally, our hierarchical framework is also an important contribution, as we model theme-specific intercepts and slopes in a specification that is relevant to the notion of factors being depicted as collections of distinct themes, recognizing that both alpha and market exposures vary across themes, a specification that is supported by the data. In particular, the Bayesian versions of this model are especially relevant in an EM data-scarce context.

Our analysis is based on a large-scale panel of equity data covering multiple EM regions over a long sample period. The dataset spans thousands of stocks and several hundred factor signals, organized into economically motivated themes, and is evaluated across four regions: the full aggregated Emerging Markets universe, the Americas, APAC, and EMEA. This regional decomposition allows us to assess both the global robustness of factor premia and meaningful cross-market heterogeneity. By combining breadth across signals with depth across time and geography, the data provide a demanding environment to evaluate the existence, stability, and economic relevance of factor alphas under alternative inferential frameworks.

Our results show that the factor premia that emerge most robustly in Emerging Markets are precisely those that occupy a central position in the asset pricing literature and also have a higher number of proposed causal mechanisms (Aghassi et al., 2023). Across regions, themes related to value, profitability, quality, profit growth, and momentum consistently display positive and economically meaningful alphas, while more peripheral themes exhibit weaker and less stable behavior. Organizing the factor zoo around economically coherent themes offers a substantially more structured and parsimonious framework than suggested by the proliferation of individual signals, with some theme-level results contradicting factor-level tests. Empirically, both alphas and market exposures vary systematically across themes, and hierarchical specifications receive strong support from the data. Finally, linking inference to out-of-sample portfolio performance shows that ignor-

ing multiple testing generally leads to inferior outcomes, while excessively conservative corrections increase risk. Methods that balance false discovery control and information pooling, such as FDR-based procedures and hierarchical Bayesian models with informative priors, offer a middle ground, delivering more stable inference and superior risk-adjusted performance in several regions. Overall, out-of-sample improvements are not universal and depend on both geography and individual preferences.

2. Literature Review

2.1. P-hacking

The asset pricing literature has been going through considerable controversies, given the widespread discovery of hundreds of factors in the context of the factor zoo (Cochrane, 2011).

The usual methods for testing factors or anomalies consist of running Fama-MacBeth 2-step regressions or building characteristics-based portfolios and then running regressions against benchmark factor models. Researchers then apply single t-tests on risk premiums or alphas in order to evaluate statistical significance, usually applying a 5% threshold (Goyal, 2012).

Harvey et al. (2016) highlight a central concern in this literature, mentioning that the high number of documented factors is caused by a p-hacking explanation caused by publications biases. Because non-discoveries and replication studies are often difficult to publish and attract limited attention, researchers are incentivized to test many specifications until statistically significant results emerge. As the number of tests increases, so does the expected number of Type I errors, causing nonexistent factors to appear statistically significant and generating false discoveries (Harvey, 2017). In fact, economics and business are among the fields with the lowest publication rates for non-significant findings (Fanelli, 2010).

Harvey et al. (2016) support this explanation by showing that, in their compilation of factor studies, the number of t-statistics between 2.0 and 2.57 is nearly identical to those between 2.57 and 3.14, with very few observations below 2.0, a pattern consistent with selective reporting. Relatedly, Harvey (2017) notes that, if anomalies were genuine, their discovery rate should decline over time, as the most obvious anomalies and factors tend to be discovered early. However, the number of published discoveries has risen sharply.

Importantly, this phenomenon is not unique to finance. Using hundreds of thousands of confidence intervals from Medline (PubMed), Barnett and Wren (2019), as processed into z-values by van Zwet and Cator (2021), document a similarly distorted distribution. Along the same lines, Ioannidis (2018) report that over 90% of published p-values are below 0.05, an outcome that is statistically implausible, with further evidence provided by Head et al. (2015).

This environment leads to systematic overestimation of effect sizes and overly narrow confidence intervals, a phenomenon known as the *winner's curse*. The publication-driven "significance filter" biases estimates upward, with the magnitude of this bias decreasing in statistical power: it is most severe when true effects are small and standard errors are large, conditions that are common in asset pricing due to low signal-to-noise ratios (Ioannidis, 2018; van Zwet and Cator, 2021).

Although only recently mainstream in asset pricing, this scientific problem is not new. Several decades ago, Rozeboom (1960) argued that statistical inference had become a procedural tool aimed at achieving a goal through null-hypothesis testing, creating a "null-hypothesis test" dogma. This dogma reduces the cognitive process of believing in a hypothesis to a binary decision based on an arbitrary threshold applied to a single observational design, thereby undermining the role of criticism in science.

Meehl (1978) expanded this critique, arguing that the dominance of significance testing provides little insight into underlying mechanisms or genuine theoretical falsification, a point made in the asset pricing context by De Prado (2023). Although Meehl (1978) focused on psychology, his arguments readily extend to asset pricing. Financial research has widely adopted Fisher's null-hypothesis framework while often neglecting Popper's view that scientific progress requires theories with falsifiable predictions. Even worse, asset pricing theories are frequently built inductively to match observational evidence (Maiti, 2019).

A central point in Meehl (1978) is that the null hypothesis is almost always false in practice. Outcomes depend on many small causal inputs that are never perfectly balanced between groups, making a zero-difference null rarely realistic (Cohen, 1994). With enough data, even tiny effects with no economic meaning will eventually reject the null, offering little in terms of Popperian refutation. Moreover, measurement error and model misspecification can easily generate statistically significant results with limited substantive meaning (Meehl, 1978; Berkson, 1942).

In an asset pricing context, consider long-short portfolios formed on a characteristic, such as book-to-market. High and low portfolios almost inevitably differ in a variety of distinct secondary risk exposures. When theories are vague, lacking specification of structural mechanisms, proper treatment of confounders and colliders, presence of falsifiable auxiliary assumptions in the Lakatosian sense (eg. trading frictions, liquidity, sample choice), or clear quantitative predictions (e.g., precise horizons, expected signs across regimes, or magnitudes), a small effect combined with sufficient data is frequently interpreted as evidence against the null of zero alpha. In addition to that, the very way we measure a factor alpha, i.e. a backtest, can in itself be full of possible measurement errors, such as survivorship and look-ahead biases, poorly estimated transaction costs and accounting restatements (De Prado, 2023; Bailey et al., 2014; Asness and Frazzini, 2013; Harvey, 2017).

Selection bias can occur at many different stages during research, with many trail-and-error attempts that go undocumented generating a "garden of forking paths" (Ioannidis, 2018; Gelman, 2015). In asset pricing, this translates to cherry-picking the time period, liquidity thresholds, factors definitions, models specifications, portfolio construction methods, data preprocessing choices and other decisions.

This gets worse with big data, which expands the number of possible analyses and the potential for false positives. Given advances in computational power, running thousands of backtests became a trivial task. Therefore, the number of reported factors is likely an under-representation of the total number of strategies tested, with authors picking the one with smallest p-value (Harvey et al., 2016; Harvey, 2017; Chordia et al., 2020). For this reason, Bailey and López de Prado (2014) argue that all trials related to a backtested strategy should be disclosed to allow estimation of its false discovery probability and propose the deflated Sharpe ratio to account for multiple testing.

In this setting, inference based on single-test procedures is invalid, as conventional significance thresholds can generate high false discovery rates (Harvey and Liu, 2019). In a single-test framework, the significance level α controls the probability of incorrectly rejecting a true null hypothesis for an individual strategy, but it fails to do so when many strategies are tested simultaneously. As a result, p-value adjustments that account for the joint occurrence of Type I errors are required.

The problem of multiple testing has been studied since the mid-twentieth century, with a variety of methods developed to control the Family Wise

Error Rate (FWER), the False Discovery Proportion (FDP), and the False Discovery Rate (FDR). Some procedures aim to control FWER, defined as the probability of at least one false discovery, including Bonferroni, Holm (1979), Hommel (1988), Hochberg (1988), and Romano and Wolf (2005). Other approaches focus on controlling the FDP or, more commonly, the FDR, defined as the expected proportion of Type I errors among rejected hypotheses, such as Benjamini and Hochberg (1995) (BH) and Benjamini and Yekutieli (2001) (BY). FDR-based methods are less conservative and therefore allow more strategies to be declared profitable as the number of tests increases, though all procedures rely on specific assumptions regarding dependence across hypotheses (Chordia et al., 2020).

Harvey et al. (2016) apply Bonferroni, Holm, and BY adjustments to a set of 316 factor t-statistics, using significance levels of 5 percent for Holm and 1 percent for BY. Because Bonferroni and Holm thresholds increase with the number of tests, the implied critical t-statistics range from about 2.0 to nearly 4.0. In contrast, BY-adjusted thresholds are stationary and converge to approximately 3.4, or 2.8 when using a 5 percent level, which the authors propose as a minimum significance threshold for asset pricing. Correspondingly, under Bonferroni and Holm FWER-focused methods, rejection rates fall below 0.1 percent, while, under BH and BY, they rise to 8.1 percent and 0.9 percent, respectively, highlighting the low power of FWER-based procedures.

Relatedly, Chordia et al. (2020) generates more than two million strategies with zero-mean alphas, varying signal informativeness and signal-to-noise ratios. Under classical hypothesis testing, between 12 and 35 percent of strategies appear significant depending on portfolio construction and factor benchmarks. FWER-based adjustments are again shown to be overly conservative, particularly when signal-to-noise ratios are low. FDR and FDP methods yield higher rejection rates, with BH being the least conservative, BY the most conservative, and Romano-Wolf intermediate.

The authors further show that BH can exhibit high variability in the false discovery proportion, especially when signals are correlated, whereas BY keeps FDP variability low at the cost of very low power. Overall, power increases with the signal-to-noise ratio, while a larger number of tested strategies mainly reduces FDP variability rather than raising rejection thresholds.

Not all authors view p-hacking as sufficient to explain the factor zoo. Chen (2021) shows that implausibly many search attempts would be required to generate the largest reported t-statistics, suggesting that very high values

(above 4.0) are likely true discoveries, a conclusion actually shared by Harvey et al. (2016) and Chordia et al. (2020). At the same time, Chen (2022) argues that mechanically raising significance thresholds is difficult to justify and may rely on overly conservative assumptions about false discoveries.

Bayesian approaches offer an alternative by incorporating prior information, inducing shrinkage that mitigates the winner's curse, and allowing theory to influence inference (van Zwet and Cator, 2021; Harvey, 2017). Harvey (2017) emphasizes the relevance of Bayesian methods in asset pricing. He argues that mechanically raising t-statistic thresholds through frequentist multiple-testing adjustments may unintentionally increase publication bias and data snooping, thereby exacerbating p-hacking, consistent with Goodhart's law. Moreover, when true effects are rare, even high-powered tests can yield a high probability of false discoveries.

Finally, null hypothesis testing suffers from structural limitations. Exact zero effects are rarely realistic, and conventional significance thresholds such as 0.05 are inherently arbitrary, with little practical distinction between results just above or below the cutoff. Bayesian practice explicitly acknowledges this arbitrariness, for example by reporting credible intervals at unconventional levels such as 89 percent (McElreath, 2018).

Jensen et al. (2023) apply both frequentist (BY) and Bayesian approaches to multiple testing in a dataset of 153 factors across 93 countries. Their Empirical Bayes framework clusters factors into 13 themes, rather than as independent signals, a view also emphasized by Aghassi et al. (2023). Factor alphas are decomposed into global, theme-level, signal-level, and idiosyncratic components, with dispersion in CAPM alphas used to calibrate empirical Bayes priors at each level. Using this approach, the authors report replication rates of 75.6 percent under BY and 82.4 percent under the empirical Bayes framework, suggesting favorable evidence for factor robustness.

To quantify publication bias, Chen and Zimmermann (2020) model the observed t-statistic as the sum of a standardized true expected return and a standardized sampling error, assuming a zero-mean normal prior for the true effect with variance 3.0^2. Using an empirical Bayes framework, they obtain an average shrinkage correction of about 10 percent, consistent with related estimates ranging from 8 to 17 percent in Harvey et al. (2016), Chen and Zimmermann (2020), and Jensen et al. (2023).

A recurring conclusion in this literature is that factors grounded in scientific theory should exhibit lower false discovery rates, as their prior probability of being true is higher (Aghassi et al., 2023; Harvey, 2017; Harvey et al.,

2016). As the asset pricing literature matures and genuine anomalies become rarer, theory plays an increasingly important role in guarding against false discoveries (Aghassi et al., 2023; Chordia et al., 2020). Bayesian frameworks naturally accommodate this distinction by allowing theory-backed factors to receive higher prior probabilities. In contrast, reliance on inductive pattern discovery without falsifiable mechanisms contributes to the credibility concerns surrounding the factor zoo (De Prado, 2023).

2.2. Emerging Markets

EM can be characterized by economies with high volatility, a transitional character, occurring in multiple spheres in the society, and smaller size, liquidity and market accessibility, covering countries in Latin America, Eastern Europe, Africa, Asia-Pacific and Middle East (Mody, 2004; MSCI, 2023).

Despite expected to account for around 60% of global GDP by 2026, such countries are underrepresented in literature, which highlights a home bias effect (Karolyi, 2016; Guimarães and Kimura, 2025). Regarding asset pricing, they display several other challenges. First of all, effective samples sizes in EM are much smaller, with many of those countries only starting to have modernized exchanges in late 1990s to beginning of the 2000s, and even so with narrow breadth. Additionally, such return series usually have more structural breaks and regime changes, more synchronous movements, lower signal-to-noise, lower convergence with international accounting standards, periods of hyperinflation, more cross-sectional concentration and possible more survivorship bias (Harvey, 1994; Morck et al., 2000). Those factors can possibly lead to biased estimates of risk premiums, especially for characteristics that are not available throughout the entire EM sample.

In this data-scarce context, hierarchical models that allow for partial pooling are of great importance, because, when estimating group-level parameters, one can better explore the data, borrowing strength across groups, helping with variance and overfitting reduction. In this context, Bayesian models can be even more relevant, given shrinkage effects with introduction of priors, that can allow better handling of high dimensionality (Gelman, 2015).

A branch of the finance literature studies the degree of international market integration using diverse methods and research designs. For example, Bekaert and Harvey (1997) show that world factors explain an increasing share of national return volatility, consistent with rising integration. Rapach et al. (2013) document strong lead-lag effects from US to other developed

markets, but not in the reverse direction. Using the December 2000 redefinition of the MSCI All Country World Index as a natural experiment, Hau (2011) further find evidence in favor of globally integrated risk pricing.

In fact, some studies replicate anomalies and factor models in EM and compare global versus local versions of asset pricing models. Zaremba and Czapkiewicz (2017) replicate 100 anomaly-based long-short portfolios in Eastern Europe, grouped into 16 categories. Hanauer and Linhart (2015) document strong size, value, and momentum premiums across four EM regions and show that local factor models outperform global ones. Similarly, Cakizi et al. (2013) find robust value effects in EM and momentum everywhere except Eastern Europe, with local models again delivering superior performance, consistent with market segmentation.

EM offer a particularly informative setting for studying multiple testing and factor inference. Smaller effective sample sizes, higher volatility, frequent structural breaks, and lower signal-to-noise ratios make statistical inference more fragile, amplifying the risks of false discoveries and overfitting, and thus the importance of partial pooling and Bayesian shrinkage, which are core pillars of our current framework. In particular, partial market segmentation implies that international evidence may provide useful prior information without fully determining local outcomes, motivating the use of global data to inform priors.

At the same time, a growing body of evidence documents the existence of factor premiums in EM, while also showing that their behavior differs from developed markets, with local factor models often outperforming global ones. This combination makes EM a demanding out-of-sample environment to assess asset pricing credibility.

3. Methods and Data

This study proposes a factor selection framework to address p-hacking concerns in the factor zoo. The framework combines (i) classical no-pooling, anomaly-by-anomaly tests with explicit frequentist control for multiple comparisons, and (ii) hierarchical, partially pooled models estimated under either frequentist or Bayesian paradigms, which introduce shrinkage across related factors. Therefore, the framework provides an approach to select characteristics for (multi-)factor portfolios. The partially pooled specification, particularly under Bayesian estimation, is especially well suited to EM, where samples are shorter and noisier.

We organize factor portfolios into economically interpretable themes, following common conventions in the literature. Even though the average correlation among factors has been documented to be low by some studies, there is both a semantic and quantitative hierarchical structure in the factor space, with a within-cluster pairwise correlation often above 0.5 (Jensen et al., 2023; Zaremba and Czapkiewicz, 2017; Aghassi et al., 2023; Chen, 2022; Green et al., 2013).

To answer our research questions, we first consider traditional frequentist no-pooled CAPM specifications in which parameters are estimated separately for each factor portfolio, applying a range of FWER and FDR frequentist multiple-comparisons adjustments, while also computing the no-adjustment case as a benchmark. Moreover, we consider hierarhical CAPM-style specifications in which returns are decomposed into theme-level components and factor-specific random deviations, allowing intercepts and market factor exposures to vary across themes. Under the Bayesian specification, both a conservative and an informative prior regimes are implemented, providing two distinct views.

Finally, to evaluate economic relevance of the competing selection rules, we embed each approach in a walk-foward expanding out-of-sample backtest: after an initial buffer period, models are fit and factors are selected, then selected factors are combined using either equal-weights or weights proportional to (posterior) information ratios, and then out-of-sample performance is tracked over the subsequent year, after which the entire procedure is re-estimated on an expanding window and repeated.

While our framework has conceptual links with Jensen et al. (2023), it differs in the multilevel specifications explored, the Bayesian framework and priors regimes employed, the variety of multiple testing controls tested, the EM focus, and the backtest strategy implementation.

3.1. Frequentist Multiple Comparisons Adjustments

3.1.1. Family-Wise Error Rate Control

In multiple hypothesis testing, the Family-Wise Error Rate (FWER) consists of the probability of making at least one False Discovery (FD):

$$FWER = P(FD \geq 1) \tag{1}$$

and FWER control at level α requires $P(FD \geq 1) \leq \alpha$.

Bonferroni Correction

Bonferroni one-step procedure rejects hypothesis H_i, for $i = 1, \ldots, N$ tests whenever associated p-value p_i:

$$p_i \leq \frac{\alpha}{N} \tag{2}$$

By the Bonferroni inequality, this guarantees strong FWER control under arbitrary dependence among tests (Holm, 1979; Hochberg and Tamhane, 1987). While simple and robust, Bonferroni is well known to be conservative.

Holm's Procedure

The Holm (Bonferroni-Holm) method is an uniformly more powerful step-down procedure. Ordering the p-values: $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(N)}$, hypothesis are sequentially rejected as long as:

$$p_{(i)} \leq \frac{\alpha}{N + 1 - i} \tag{3}$$

Rejection stops at the first failure, and all remaining hypotheses are retained. Holm's method can be interpreted as the closed-testing procedure obtained by applying Bonferroni tests to all intersection hypotheses, which explains its strong FWER control without dependence assumptions (Holm, 1979; Hommel, 1988).

Hochberg and Hommel procedures

Building on Simes' inequality (Simes, 1986), Hochberg (Hochberg, 1988) proposes a step-up FWER-controlling procedure that is more powerful than Holm under independence or certain positive dependence structures.

Given ordered $p_{(1)} \leq \cdots \leq p_{(N)}$, the Hochberg procedure rejects hypotheses $H_{(1)}, \ldots, H_{(k)}$, where:

$$k \;=\; \max \left\{ i \in \{1, \ldots, N\} : p_{(i)} \leq \frac{\alpha}{N + 1 - i} \right\}.$$

If no such i exists, no hypothesis is rejected. Unlike step-down procedures, Hochberg's method starts from the largest p-values and moves upward, yielding greater power when its dependence assumptions hold.

Hommel's procedure (Hommel, 1988) further exploits the closure principle by using the full family of Simes-type intersection tests. Let $p_{(1)} \leq \cdots \leq p_{(N)}$

denote the ordered p-values. Define

$$j^* \;=\; \max\left\{ j \in \{1,\dots,N\} : p_{(N-j+k)} > \frac{k\alpha}{j} \;\text{ for all } k = 1,\dots,j \right\}.$$

If no such j^* exists, all hypotheses are rejected. Otherwise, the final rejection rule is

$$\text{reject all } H_{(i)} \text{ such that } p_{(i)} \leq \frac{\alpha}{j^*}.$$

Both procedures control the FWER under independence and dominate Holm in power, while remaining conservative in low signal-to-noise settings with many hypotheses (Chordia et al., 2020). Hommel's method uniformly dominates Hochberg in power, additionally retaining strong control under a wider range of dependence structures, but is computationally more demanding.

Overall, FWER-controlling procedures provide strong error guarantees but typically suffer from low power when the number of tests is large.

3.1.2. False Discovery Rate Control

False Discovery Rate (FDR) control targets the expected proportion of false rejections among all rejections. Let FD and TD denote the number of false and true discoveries, respectively, and define the false discovery proportion (FDP) as

$$FDP \;=\; \begin{cases} \dfrac{FD}{FD+TD}, & \text{if } FD+TD>0, \\ 0, & \text{if } FD+TD=0. \end{cases} \tag{4}$$

The false discovery rate (FDR) is the expectation of the FDP:

$$FDR \;=\; \mathbb{E}\big[FDP\big]. \tag{5}$$

The FDR is $\mathbb{E}[FDP]$. Controlling FDR is less stringent than FWER control and typically yields substantially higher power when many hypotheses are tested (Benjamini and Yekutieli, 2001).

Benjamini-Hochberg Procedure

The BH step-up procedure (Benjamini and Hochberg, 1995) orders p-values and rejects all hypothesis with:

$$p_{(i)} \leq \frac{i}{N}\alpha \tag{6}$$

for the largest such i. Under independence, BH controls the FDR at level α, and it remains valid under a broad class of positive dependence structures (Benjamini and Yekutieli, 2001). Compared to FWER methods, BH rejects more hypotheses and exhibits substantially higher power.

Benjamini-Yekutieli Procedure

To ensure FDR control under arbitrary dependence, Benjamini and Yekutieli (2001) propose a conservative modification of BH by replacing α with $\frac{\alpha}{c(N)}$, where $c(N)$ is the harmonic correction:

$$c(N) = \sum_{i=1}^{N} \frac{1}{i}, \tag{7}$$

and satisfies $c(N) > 1$ for $N \geq 2$ ($c(N) \approx \log N + \gamma + \frac{1}{2N}$, where γ is the Euler-Mascheroni constant). The resulting thresholds:

$$p_{(i)} \leq \frac{i}{N \cdot c(N)}\alpha \tag{8}$$

guarantee FDR control under any dependence structure but can be substantially conservative for large N. When positive dependence (e.g., PRDS or MTP2) is plausible, the original BH procedure is preferred for power.

3.2. Hierarchical Models

Hierarchical (multilevel or mixed-effects) models interpolate between two extremes: complete pooling and no pooling. Complete pooling ignores clustering and treats all observations as independent, while no pooling estimates separate models for each factor, risking overfitting, particularly for short samples, and implicitly assuming that related factors contain no shared information. This assumption is implausible in asset pricing, as factors within the same theme (e.g., valuation multiples or momentum signals) often share constituents and exhibit substantial within-theme correlation (Aghassi et al., 2023).

Partial pooling relaxes both extremes by allowing factor-level heterogeneity while borrowing strength across related factors within the same theme. This structure captures semantic and statistical dependencies across factors

and avoids discarding informative cross-factor variation, improving estimation efficiency and robustness (Johnson et al., 2022). Accordingly, we adopt a nested structure with monthly factor returns grouped within themes.

3.2.1. Model Specification

Let i index factors, c index themes, and t index time, with factor i belonging to theme $c(i)$. We estimate the hierarchical return model

$$r_{i,t} = \alpha_{c(i)} + \beta_{c(i)}\, mrkt_t + a_i + b_i\, mrkt_t + \varepsilon_{i,t}, \tag{9}$$

where $r_{i,t}$ denotes excess factor returns and $mrkt_t$ is the market factor. The parameters $\alpha_{c(i)}$ and $\beta_{c(i)}$ are theme-level fixed effects for average abnormal returns and market exposure given each theme.

Factor-level deviations are modeled as random effects:

$$\begin{pmatrix} a_i \\ b_i \end{pmatrix} \sim \mathcal{N}(\mathbf{0}, \Sigma)\,, \qquad \varepsilon_{i,t} \sim \mathcal{N}(0, \sigma^2), \tag{10}$$

where a_i and b_i represent factor-specific deviations in intercepts and slopes. The covariance matrix Σ allows for intercept–slope correlation, and $\varepsilon_{i,t}$ captures idiosyncratic noise.

This structure implies that factor alphas decompose into a theme-level component plus a factor-specific deviation, while allowing heterogeneous market exposures across factors[2]. Estimation therefore borrows information across factors within themes while preserving idiosyncratic variation.

We assess robustness using two reduced specifications. First, we impose a common theme-invariant market loading,

$$r_{i,t} = \alpha_{c(i)} + \beta\, mrkt_t + a_i + b_i\, mrkt_t + \varepsilon_{i,t}, \tag{11}$$

and second, a fully factor-level fixed-effects model,

$$r_{i,t} = \alpha + \beta\, mrkt_t + a_i + b_i\, mrkt_t + \varepsilon_{i,t}. \tag{12}$$

[2]while our data consists of long-short zero-net-investment portfolios, they are not zero-beta (Jensen et al., 2023)

3.2.2. Estimation

We estimate the model using both Bayesian and frequentist approaches. Bayesian estimation is conducted via full posterior inference and naturally induces partial pooling, shrinking noisy factor-level estimates toward theme-level means and stabilizing inference in short and volatile EM samples. Frequentist estimates, which are additionally used to inform prior construction, are produced by a Gaussian linear mixed-effects model via restricted maximum likelihood (REML), which yields less biased variance component estimates in finite samples. Fixed effects are inferred using Wald-type statistics with Satterthwaite-approximated degrees of freedom to account for uncertainty in estimated variance components (Kuznetsova et al., 2017).

Prior regimes

We consider two prior strategies.

(i) Weakly informative (skeptical) shrinkage priors.

Theme-level alphas and market loadings follow zero-mean normal priors, while variance components follow half-t distributions,

$$\alpha_c \sim \mathcal{N}(0, s_\alpha^2), \qquad \beta \sim \mathcal{N}(0, s_\beta^2), \tag{13}$$

$$\mathrm{sd}(a_i) \sim \text{half-}t(\nu, 0, s_a), \qquad \mathrm{sd}(b_i) \sim \text{half-}t(\nu, 0, s_b), \tag{14}$$

$$\sigma \sim \text{half-}t(\nu, 0, s_r), \qquad \mathrm{Cor}(a_i, b_i) \sim \mathrm{LKJ}(\eta), \tag{15}$$

with scales calibrated to monthly returns of long–short factor portfolios. [3]

(ii) Informative priors derived from global data.

A growing literature documents that several factor anomalies exhibit positive and persistent alphas across regions, periods, and even asset classes (Ilmanen et al., 2021; Aghassi et al., 2023), while also emphasizing the presence of regional idiosyncrasies driven by differences in risk exposures, market structure, and behavioral mechanisms (Zaremba and Czapkiewicz, 2017; Asness, 2011; Jacobs and Müller, 2020), as well as partial integration between local and global factor markets (Fama and French, 2017; Hanauer and Lauterbach, 2019). Taken together, this evidence suggests that global factor premia provide informative anchors for EM, but should be allowed to adjust to region-specific dynamics, particularly given shorter samples, higher

[3] We set $s_\alpha = 0.075$ (monthly), $s_\beta = 0.05$, $\nu = 5$, $s_a = 0.10$, $s_b = 0.05$, $s_r = 3.0$, and $\eta = 2$ for the LKJ prior. These values are weakly informative relative to the empirical dispersion of factor returns.

concentration, and greater macroeconomic volatility in EM (Harvey, 1995).

Accordingly, we first estimate the hierarchical model in (9) on a large global dataset and map the resulting estimates into informative priors for the EM analysis. Concretely,

$$\alpha_{c(i)} \sim \mathcal{N}\big(\hat{\alpha}^{G}_{c(i)},\ \mathrm{SE}(\hat{\alpha}^{G}_{c(i)})\big)\,, \qquad \beta_{c(i)} \sim \mathcal{N}\Big(\hat{\beta}^{G}_{c(i)},\ \mathrm{SE}(\hat{\beta}^{G}_{c(i)})\Big)\,, \tag{16}$$

with variance priors scaled to global dispersion estimates.

Posterior summarization and inference

Posterior draws $\{\tilde{\theta}^{(s)}\}_{s=1}^{S}$ are summarized by medians and 89% credible intervals. Factor-level parameters are constructed as

$$\alpha_i^{(s)} = \alpha_{c(i)}^{(s)} + a_i^{(s)}, \qquad \beta_i^{(s)} = \beta_{c(i)}^{(s)} + b_i^{(s)}.$$

We report posterior probabilities of positive effects,

$$\mathrm{pd}^{\mathrm{B}}_{\alpha_i} = \Pr(\alpha_i > 0 \mid \text{data}),$$

alongside posterior means, standard deviations, and signal-to-noise ratios.

Bayesian inference complements frequentist t-statistics by quantifying uncertainty through full posterior distributions.

We additionally compute Bayes factors,

$$\mathrm{BF}_{10} = \frac{\mathbf{P}(\mathcal{D} \mid M_1)}{\mathbf{P}(\mathcal{D} \mid M_0)},$$

to compare models with and without non-zero alphas. Posterior probabilities, Bayes factors, and frequentist statistics together provide a nuanced assessment of economic and statistical significance (Harvey, 2019).

3.3. Walk-Forward Expanding Backtests

To assess the economic implications of alternative factor-selection methodologies, we implement expanding walk-forward backtests that map statistical selection rules into portfolio choices and evaluate their out-of-sample performance.

3.3.1. Backtest Scheme

We employ an expanding-window walk-forward design that preserves the time ordering of returns (Tashman, 2000; Schnaubelt, 2019; Bergmeir et al., 2018). Let $t = 1, \ldots, T$ index monthly observations and let $\mathcal{R} = \{t_1, \ldots, t_K\}$ denote refit dates. The first refit date t_1 is chosen such that an initial training window of $N_0 = 120$ months is available.

At each refit date $t_k \in \mathcal{R}$, model parameters are re-estimated using all information available up to t_k, including external data used to construct informative priors. For Bayesian specifications, posterior draws are obtained and factor eligibility is re-evaluated under the corresponding selection rule.

Formally, the training sample at refit k is

$$\mathcal{T}_{\text{train}}^{(k)} = \{1, \ldots, t_k\},$$

while the out-of-sample holding period is

$$\mathcal{T}_{\text{test}}^{(k)} = \{t_k, \ldots, t_{k+H} - 1\}.$$

During $\mathcal{T}_{\text{test}}^{(k)}$, the set of selected factors is held fixed and portfolio returns are recorded monthly. Refits occur only at a subset of calendar months (e.g., annually), generating a holding period of H months between refits. This approach substantially reduces computational cost (Gu et al., 2020).

All decisions at t_k are based exclusively on $\mathcal{T}_{\text{train}}^{(k)}$, ensuring a strictly out-of-sample evaluation and eliminating forward-looking bias.

3.3.2. Selection Strategies

We consider two classes of factor-selection strategies.

(i) No-pooled frequentist strategies. Let $\mathrm{f} \in \{1, \ldots, 6\}$ index the frequentist multiple-testing adjustments described in Section 3.1, including the unadjusted case. Under strategy f, a factor i is selected at refit k to the set $\mathcal{F}_f^{(k)}$ if its associated p-value is both statistically significant and positive,

$$i \in \mathcal{F}_{\mathrm{f}}^{(k)} \iff p\text{-value}_{\mathrm{f}}^{(k)} \leq 0.05. \tag{17}$$

(ii) Partial-pooled Bayesian strategies. Let $\mathrm{b} \in \{1, 2\}$ index Bayesian hierarchical specifications under alternative prior regimes. A factor i is selected at refit k if its posterior probability of a positive alpha exceeds 95%,

$$i \in \mathcal{B}_{\mathrm{b}}^{(k)} \iff \mathrm{pd}_{\alpha_i,\mathrm{b}}^{\mathrm{B},(k)} \geq 0.95. \tag{18}$$

3.3.3. Portfolio Construction

Selected factors are combined into multi-factor portfolios using a portfolio-blending approach, in which factor portfolios are treated as individual assets. This design promotes diversification at the expense of attenuated signal exposure.

We consider three weighting schemes.

(i) Equal weights:

$$w_{i,\mathcal{S}}^{ew,(k)} = \frac{1}{|\mathcal{S}^{(k)}|}, \tag{19}$$

where $||$ denotes the number of selected factors of a generic selection set $\mathcal{S}^{(k)}$. This scheme treats signal existence as sufficient, abstracting from effect magnitude.

(ii) Weights proportional to no-pooled t-statistics:

$$w_{i,\mathcal{F}}^{(k)} = \frac{t_{\alpha_i}}{\sum_{i\in\mathcal{F}^{(k)}} t_{\alpha_i}}, \tag{20}$$

which assigns larger weights to factors with higher signal-to-noise ratios under frequentist estimation.

(iii) Weights proportional to partial-pooled Bayesian t-statistics:

$$w_{i,\mathcal{B}}^{(k)} = \frac{t_{\alpha_i}^{\mathrm{B}}}{\sum_{i\in\mathcal{B}^{(k)}} t_{\alpha_i}^{\mathrm{B}}}, \tag{21}$$

where Bayesian shrinkage moderates extreme weights through hierarchical pooling.

3.4. Dataset

We use the Global Factor Data of Jensen et al. (2023), which covers 406 characteristics aggregated into 153 factors, providing broad coverage of the factor zoo. For each characteristic, country, and month, stocks are sorted into terciles using country-specific breakpoints based on non-micro stocks. Factor portfolios are constructed as capped value-weighted returns, with weights proportional to market equity and winsorized at the NYSE 80th percentile. Excess factor returns are defined as high-minus-low tercile returns. Portfolio construction requires at least five stocks per tercile and a minimum of 60 monthly observations per country–factor, with accounting variables updated using the most recent available information.

We focus on the EM region for estimation and backtesting, while the World sample is used exclusively to construct informative priors. From the Global Factor Data, we import a country-level panel covering all available countries and the EM regional panel provided by Jensen et al. (2023). Using MSCI market classifications (MSCI Inc., 2025), we restrict the country-level panel to EM countries and define three custom EM regions: Americas, APAC, and EMEA. Country-level factor returns are aggregated into regional panels via equal-weighted averages across countries, preserving the original within-country value-weighting and capping rules.

Regional factor panels are merged with market and risk-free proxies. For EM regions, we use the MSCI EM index as the market proxy, while for the global sample we use the MSCI ACWI index, extending its history with the S&P 500 when necessary. Excess market returns are computed using the U.S. Treasury bill rate. All series are merged at a monthly frequency and expressed in percentage terms.

We impose minimum data-coverage requirements at the factor–region level by tracking start dates, end dates, and available observations. Only factors with at least 240 monthly observations are retained, ensuring sufficient time-series depth for estimation and backtesting. This filter removes 17 region–factor combinations in the custom regions, primarily due to late-starting factors. An analogous summary is constructed for the JKP EM region.

Finally, we convert both the JKP EM region and each custom region into balanced factor panels. For each custom region, we restrict attention to factors passing the coverage filter and define a common sample window by intersecting factor-specific start and end dates. Remaining missing values are imputed using the cross-sectional median across factors at each date. Imputation is minimal, affecting less than 0.06% of observations. The same procedure is applied to the JKP EM region, where no imputation is required.

4. Results

Table 1 summarizes the time span and overall panel length for each of the three custom regions.

In the following subsections, we provide results framed in topics.

Table 1: Time span and length of emerging market factor panels

	Sample period		
Region	Start date	End date	Total months
APAC	2003-02-28	2024-12-31	263
Americas	2004-09-30	2024-12-31	244
EMEA	2004-08-31	2024-12-31	245
All emerging countries	2004-09-30	2024-12-31	244

This table reports the sample period and total number of monthly observations used to construct the regional factor panels for emerging markets. "All emerging countries" refers to the JKP emerging region aggregated across all emerging economies.

4.1. Existence and Magnitude of Alpha

4.1.1. Frequentist View

Table 2 reports the share of factors significant at the 5% level. Without multiple-testing correction, significance rates are high, although lower than figures for developed markets: about 51% for all EM, 52% for APAC, 24% for Americas, and 40% for EMEA. These figures are mechanically inflated by the large number of tests and the accumulation of Type I errors. The result for all EM is close to the replication rate reported by (Jensen et al., 2023).

Controlling FWER through Bonferroni, Holm, Hochberg, or Hommel sharply reduces significance to roughly 13-17% for all EM and APAC, 4% for Americas, and 9% for EMEA. These procedures are intentionally conservative, as they limit the probability of any false discovery across the full set of tests.

False discovery rate (FDR) methods lie between the unadjusted and FWER-corrected cases. BH identifies about 39% of factors as significant in all EM, 47% in APAC, 10% in Americas, and 27% in EMEA, while the more conservative BY yields 24%, 26%, 4%, and 14%, respectively. By controlling the expected share of false rejections, FDR methods allow more discoveries at the cost of tolerating some false positives.

Across all corrections, the Americas, which is the region with fewest countries, consistently display fewer significant factors, with APAC being the region with most significant factors.

Theme-level results by region, reported in Appendix A, point to a common pattern across EM. While several themes exhibit high raw discovery

Table 2: Proportion of significant factors under multiple-testing adjustments

Adjustment method	All emerging countries	APAC	Americas	EMEA
None	0.510	0.523	0.238	0.397
Bonferroni	0.150	0.134	0.041	0.089
Holm	0.163	0.154	0.041	0.089
Hochberg	0.163	0.154	0.041	0.089
Hommel	0.170	0.168	0.041	0.089
BH	0.388	0.470	0.095	0.274
BY	0.238	0.260	0.041	0.144

This table reports the fraction of characteristics classified as eligible (significant) under different multiple-testing correction procedures. "All emerging countries" refers to the JKP emerging markets panel.

rates, multiple-testing adjustments sharply reduce significance, indicating that some of the apparent factor richness reflects multiplicity rather than robust economic effects. Across regions, persistent evidence concentrates in a narrow set of themes, most notably momentum and value, followed by profitability, quality and profit growth, and to a lesser extent size, whereas the majority of themes prove highly sensitive to correction choice.

Regional heterogeneity is nonetheless present. APAC displays the richest and most resilient factor structure, with value, momentum, profitability, and profit growth retaining non-trivial shares of significant signals even under stringent corrections. The Americas cluster is markedly weaker, with robust evidence largely confined to momentum and value after adjustment. EMEA occupies an intermediate position, where momentum, value, profitability, and profit growth remain comparatively resilient, while most other themes lose significance once multiplicity is controlled.

Figure 1 shows that the choice of multiple-testing correction has a much larger impact on inference than geography itself. Unadjusted p-values cluster near zero across all regions, whereas FWER procedures (Bonferroni and Holm) shift most adjusted p-values toward one, effectively eliminating discoveries. Hochberg and Hommel yield intermediate outcomes, while the BH FDR procedure remains substantially less conservative, preserving a sizable lower tail of significant signals. BY lies between BH and FWER. Overall, statistical conclusions are highly sensitive to the correction method, with regional differences playing a secondary role.

Figure 2 complements these findings by showing that only a small subset

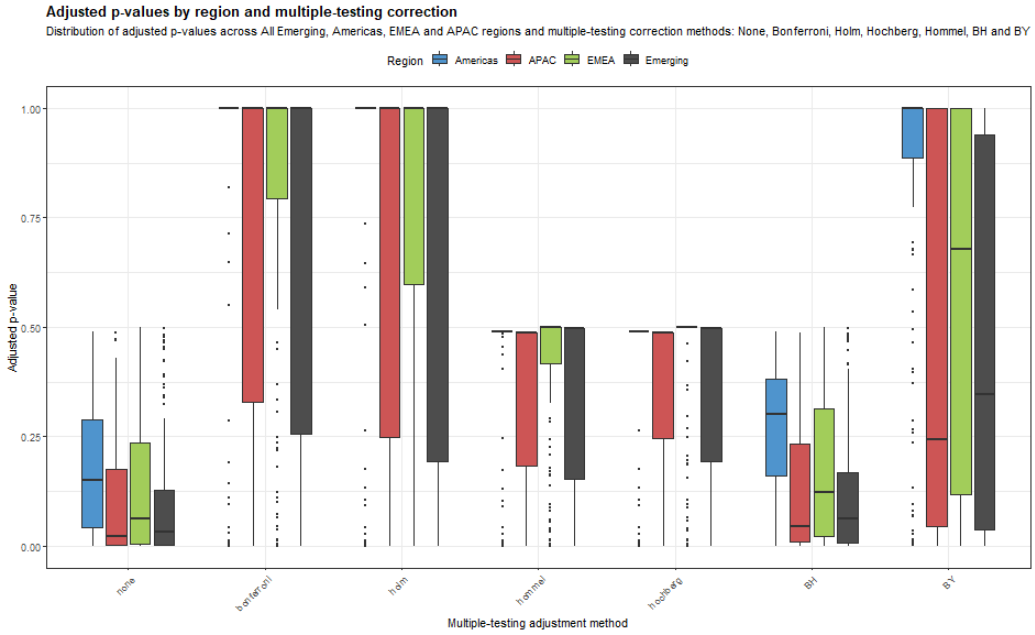

Figure 1: Adjusted p-values by correction method and region. The figure reports boxplots of adjusted p-values for all signals, grouped on the horizontal axis by multiple-testing adjustment method (None, Bonferroni, Holm, Hommel, Hochberg, BH, BY). Within each method, separate boxplots are shown for Emerging, Americas, EMEA and APAC in different colours, summarising the cross-sectional distribution of adjusted p-values across signals in each region.

of themes delivers consistently positive abnormal performance. Momentum and value stand out across all regions, with positive mean and median alpha t-statistics, often exceeding conventional significance thresholds, particularly in APAC and EMEA. Profitability and profit growth also perform well outside the Americas, while most remaining themes cluster around zero or exhibit strong regional dependence. Taken together, the evidence indicates that economically meaningful factor premia are concentrated in a few core themes, and that their apparent breadth in raw tests largely disappears once multiple testing is properly accounted for.

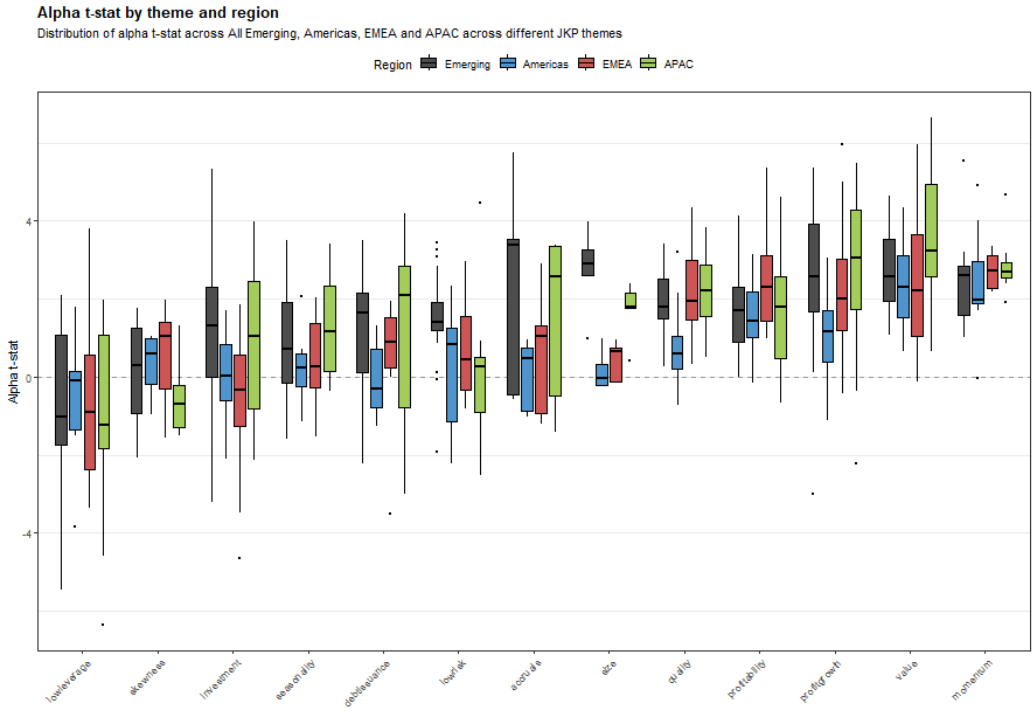

Figure 2: Alpha t-statistics by theme and region. The figure shows boxplots of the alpha t-statistic for all factors, grouped by investment theme on the horizontal axis and by region (Emerging, Americas, EMEA, APAC) in different colours. The horizontal dashed line marks zero.

4.1.2. Bayesian view

When applying the Bayesian models in (9), the analysis focuses on the posterior distributions of the model parameters. Posterior draws are obtained

using Hamiltonian Monte Carlo with 4 parallel chains and 2,000 iterations per chain, discarding the first 1,000 iterations of each chain as warm-up, with no thinning applied. Resulting Markov chain Monte Carlo output is reshaped into a single data set, with each row representing one posterior draw and each column a model parameter. Convergence is satisfactory across models, with most parameters displaying $\hat{R} \approx 1.00$ and large effective sample sizes.

Unlike the frequentist approach, which relies on p-values to declare "no" versus "some" effect, inference is based on the probability of direction. Using $\text{pd}^{\text{B}}_{\alpha_i} \geq 0.95$ as the selection criterion, Table 3 reports the proportion of factors for which more than 95% of posterior draws imply a positive α_i.

Table 3: Proportion of factors with $\Pr(\alpha > 0 \mid \text{data}) > 0.95$

Prior specification	All emerging countries	APAC	Americas	EMEA
Conservative	0.551	0.530	0.299	0.390
Informative	0.626	0.557	0.524	0.466

This table reports the fraction of characteristics classified as eligible based on their posterior alpha, using a probability of direction threshold of 95%, that is, $\Pr(\alpha_i > 0 \mid \text{data}) > 0.95$.

Under informative priors, substantially more factors are selected than under any frequentist adjustment, which is expected, but also relative to the no-adjustment case, particularly in the Americas and EMEA. Under conservative priors, selection rates are broadly comparable to the unadjusted no-pooling frequentist case and remain higher than under FWER or FDR corrections. This contrast is especially pronounced in the Americas.

Figure 3 displays the distribution of the posterior alpha signal-to-noise ratio, $t^{\text{B}}_{\alpha_i}$, defined as the posterior mean of individual alpha divided by its posterior standard deviation. Results are grouped by theme and region, with the dashed horizontal line at zero indicating the sign of the posterior mass.

Even under the conservative prior regime, which strongly shrinks alphas toward zero, a clear ranking of themes emerges. Value and momentum are the themes whose factors show the strongest and most stable positive $t^{\text{B}}_{\alpha_i}$ across all regions, while profitability, profit growth and quality factors also remain clearly positive, particularly in APAC and EMEA. Size and low risk have mildly positive factors, especially for All EM, whereas accruals, investment, debt issuance, seasonality and skewness have higher concentration near zero and show greater regional sensitivity. Low leverage stands out as the only

theme whose factors show consistently negative $t^{\mathrm{B}}_{\alpha_i}$ across regions.

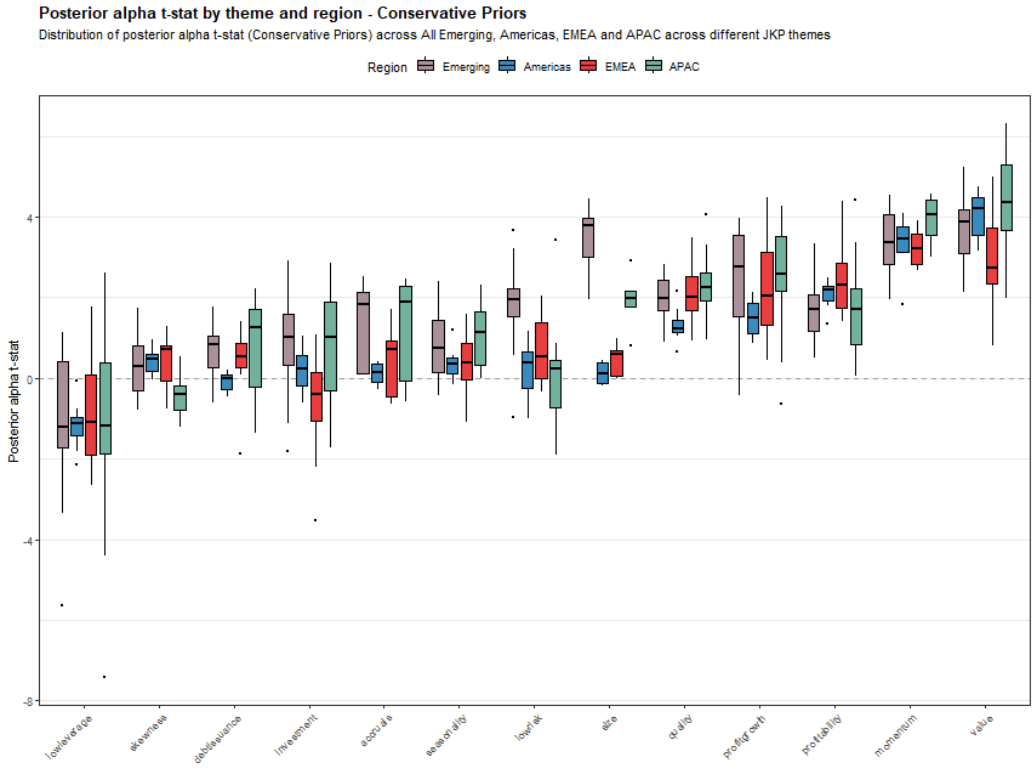

Figure 3: Distribution of posterior $t^{\mathrm{B}}_{\alpha_i}$ by JKP theme and region under the conservative prior regime. Boxplots show, for each theme (horizontal axis) and region (Emerging, Americas, EMEA, APAC), the distribution of characteristic-level $t^{\mathrm{B}}_{\alpha_i}$ (posterior mean alpha divided by its posterior standard deviation). The horizontal dashed line marks $t^{\mathrm{B}}_{\alpha_i} = 0$.

Figure 4 shows that informative priors strongly reinforce the role of the classical value and momentum themes across regions. In the full EM panel, mean posterior $t^{\mathrm{B}}_{\alpha_i}$ values are around 4 for value and 4.2 for momentum, with medians above 4 and tight interquartile ranges. Similar patterns hold regionally, with mean $t^{\mathrm{B}}_{\alpha_i}$ between roughly 3.3 and 4.7 in all regions. Profitability, profit growth and quality display uniformly positive posterior $t^{\mathrm{B}}_{\alpha_i}$ across regions, with medians between 2 and 3, indicating robust but weaker premia than value and momentum.

Size and accruals also becomes clearly positive, notably in All EM and APAC. In fact, for All EM, skewness, debt issuance, seasonality and, in special, low risk, have posterior distributions that lie entirely in positive

territory.

Investment, accruals, debt issuance and skewness shift closer to the profitable side under informative priors. In All EM and APAC, these themes reach mean $t^{\mathrm{B}}_{\alpha_i}$ close to or above 1, and, even in the Americas, several exceed 1.5. Relative to conservative priors, informative priors shift many posterior $t^{\mathrm{B}}_{\alpha_i}$ upward and slightly reduce dispersion, implying a more diversified factor selection.

Low leverage remains negative or near zero in all regions, while low risk, with the exception of All EM, shows only modest positive values.

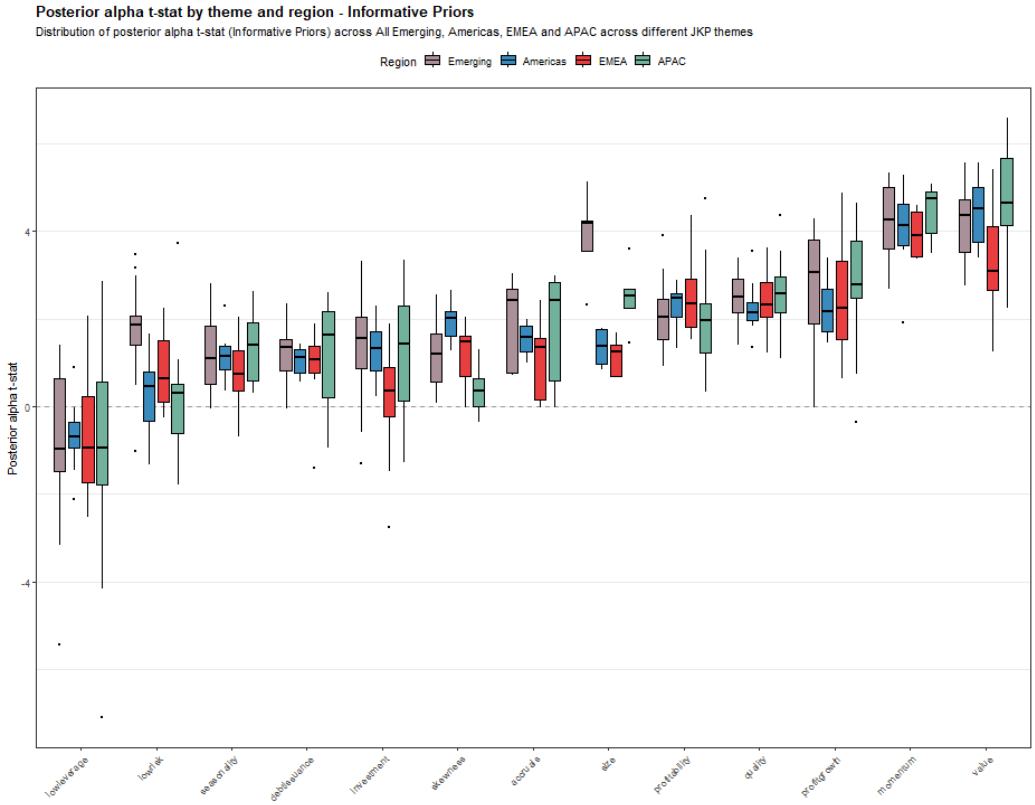

Figure 4: Distribution of posterior $t^{\mathrm{B}}_{\alpha_i}$ by JKP theme and region under the informative prior regime. Boxplots show, for each theme (horizontal axis) and region (Emerging, Americas, EMEA, APAC), the distribution of characteristic-level posterior $t^{\mathrm{B}}_{\alpha_i}$ (posterior mean alpha divided by its posterior standard deviation). The horizontal dashed line marks $t^{\mathrm{B}}_{\alpha_i} = 0$.

Additional Bayesian diagnostics are reported in Appendix B, that also

report theme-level posterior alphas. Those largely confirm the individual-level patterns shown in Figures 3–4 for momentum, value, profitability, profit growth and quality, which remain the most robust themes, with strong posterior support after aggregation.

In contrast, some themes, such as accruals, debt issuance, investment, low risk, size, skewness and seasonality, exhibit discrepancies between individual- and theme-level inference, where pooling and shrinkage generally reveal a modest positive average effect despite weak individual signals. Low leverage remains uniformly weak, with no evidence of a positive theme-level alpha across regions or prior regimes.

4.2. Hierarchical Structure

The second question is: "*do long-short factors really have a theme-based hierarchical structure with distinct parameters?*". In other words, is (9) the right theoretical framework to think about asset pricing anomalies in EM?

4.2.1. Frequentist view

Figure 5 decomposes average theme returns into alpha, the market component $\beta \times \mathbb{E}[R_M]$, and the idiosyncratic residual, for all emerging markets and by region.

Across all EM, most, but not all, themes exhibit positive alphas, with momentum, value, size, quality, profitability and profit growth having the largest average alphas, while low leverage is the only theme with a clearly negative alpha. Albeit with some particularities, patterns are broadly consistent across regions.

Being dollar-neutral long-short factors, market components are smaller than the alpha and residual terms. However, betas differ markedly across themes, and, in special, low-risk, quality, momentum, profitability and low leverage tend to be defensive, often with betas in the range -0.05 to -0.20, while others are closer to market-neutral or mildly pro-cyclical. Although statistical significance cannot be assessed from means alone, the heterogeneity suggests modelling theme-specific betas rather than imposing a common market exposure.

Figure 5: Return decomposition of theme portfolios into alpha, systematic market component ($\beta \times \mathbb{E}[R_M]$) and idiosyncratic residuals. The four panels report results for all emerging markets, APAC, Americas and EMEA, respectively.

By fitting a frequentist version of the hierarchical model in (9), we estimate theme-level alphas $\hat{\alpha}^F_{c(i)}$ and betas $\hat{\beta}^F_{c(i)}$ for each region[4]. Figure 6 shows that the resulting alpha t-statistics display remarkably similar patterns across regions.

Momentum and value consistently deliver the largest and most precisely estimated alphas, with t-statistics around 6-10, followed by quality, profitability and profit growth, with t-statistics typically between 3 and 6. It is noticeable how quality and profit growth are weaker in Americas when compared with other regions.

In contrast, remaining theme alpha t-statistics are broadly smaller than the 2.0 threshold among all regions, albeit size, low risk and investment show stronger effects for the consolidated EM cluster and APAC. Notably, low leverage displays significant negative theme alphas in all regions.

As for $\hat{\beta}^F_{c(i)}$, Figure 7 reports that the low-risk theme exhibits strongly negative and highly significant loadings on the market proxy (t-stat approx. -10 to -12), confirming that it captures portfolios with substantially lower market exposure than the universe. Investment, momentum and quality also display negative and statistically significant betas in most regions. In contrast, the beta estimates for accruals, debt issuance, skewness and seasonality are generally small and statistically indistinguishable from zero. Taken together, these results suggest that CAPM loadings do vary across themes.

Table 4 reports BH adjusted p-values for theme-specific intercepts and market factor loadings from the frequentist hierarchical CAPM across regions. Underlined entries indicate significance at the 5% level using raw (unadjusted) p-values, while bold entries denote p-values that remain significant after BH correction.

Across regions, most conclusions persist after multiple correction, with intercepts providing strong evidence of abnormal returns in the momentum, profitability, profit growth, quality and value themes. Market-factor loadings are significant for only a small subset of themes, most notably low risk, momentum, quality and, in some cases, profitability, which exhibit significant negative and countercyclical market factor exposure estimates.

To assess model specification choice, we test whether allowing for theme-specific alphas and theme-specific market betas improves the fit of the hier-

[4]Note that, differently from the last section, in which we focused on estimating alpha for individual factors, we are now interested in theme-level parameters.

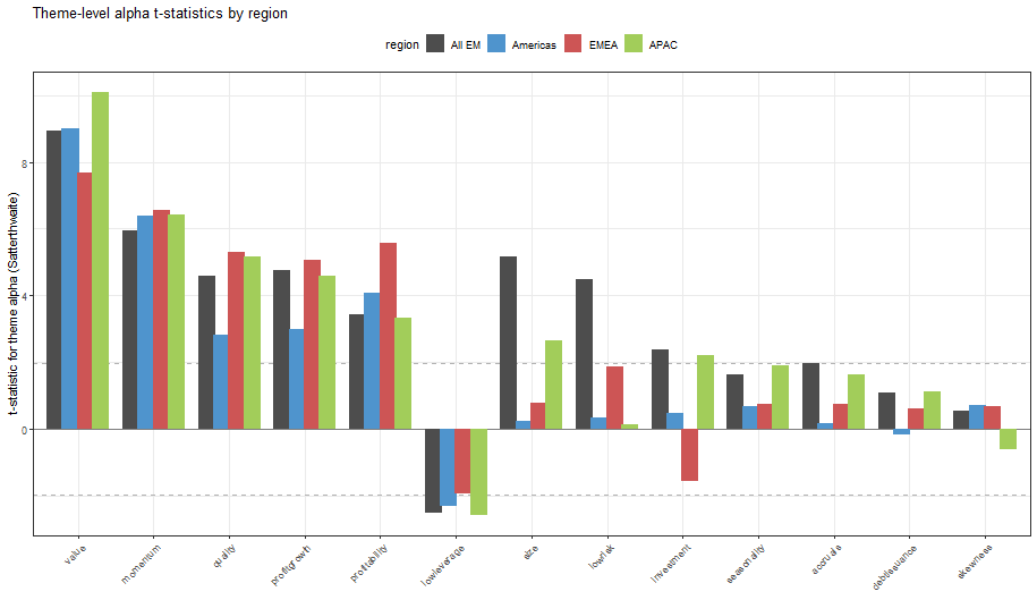

Figure 6: Theme-level alpha t-statistics $t_{\alpha_{c(i)}}$ across regions based on the frequentist hierarchical specification in Eq. 9. Bars report Satterthwaite-adjusted t-statistics for the fixed effects associated with each theme-level intercept (alphas), estimated separately for All Emerging Markets, Americas, EMEA, and APAC. The horizontal dashed lines denote the two-sided 5% critical values ($t = \pm 1.96$). Colours follow the regional grouping used in the analysis.

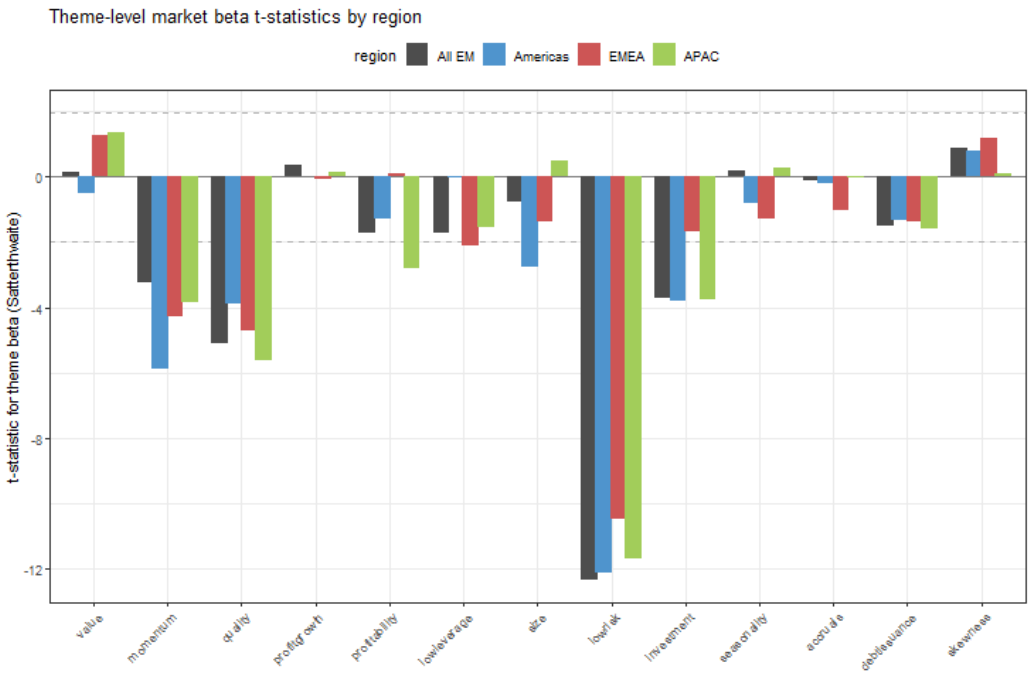

Figure 7: Theme-level beta t-statistics $t_{\beta_{c(i)}}$ across regions based on the frequentist hierarchical specification in Eq. 9. Bars report Satterthwaite-adjusted t-statistics for the fixed effects associated with each theme-level intercept (beta), estimated separately for All Emerging Markets, Americas, EMEA, and APAC. The horizontal dashed lines denote the two-sided 5% critical values ($t = \pm 1.96$). Colours follow the regional grouping used in the analysis.

Table 4: Benjamini-Hochberg adjusted p-values for theme-specific alphas and betas across regions

Theme	**BH-adjusted p-values**							
	All EM		Americas		EMEA		APAC	
	p_α^{BH}	p_β^{BH}	p_α^{BH}	p_β^{BH}	p_α^{BH}	p_β^{BH}	p_α^{BH}	p_β^{BH}
Accruals	0.073	0.897	0.941	0.941	0.530	0.415	0.181	0.980
Debt Issuance	0.353	0.171	0.941	0.388	0.583	0.301	0.364	0.181
Investment	**0.029**	**0.00044**	0.876	**0.00085**	0.247	0.211	0.058	**0.00079**
Low Leverage	**0.020**	0.124	0.052	0.992	0.139	0.107	**0.025**	0.187
Low Risk	$\mathbf{1.8 \times 10^{-5}}$	$\mathbf{< 10^{-24}}$	0.941	$\mathbf{< 10^{-21}}$	0.148	$\mathbf{< 10^{-17}}$	0.941	$\mathbf{< 10^{-20}}$
Momentum	$\mathbf{4.2 \times 10^{-8}}$	**0.00217**	$\mathbf{2.1 \times 10^{-8}}$	$\mathbf{2.2 \times 10^{-7}}$	$\mathbf{9.7 \times 10^{-9}}$	$\mathbf{1.2 \times 10^{-4}}$	$\mathbf{1.8 \times 10^{-8}}$	**0.00072**
Profitability	**0.00118**	0.124	**0.00039**	0.388	$\mathbf{9.0 \times 10^{-7}}$	0.944	**0.00324**	**0.0143**
Profit Growth	$\mathbf{7.4 \times 10^{-6}}$	0.779	**0.0106**	0.992	$\mathbf{5.4 \times 10^{-6}}$	0.944	$\mathbf{4.2 \times 10^{-5}}$	0.941
Quality	$\mathbf{1.4 \times 10^{-5}}$	$\mathbf{1.7 \times 10^{-6}}$	**0.0155**	**0.00072**	$\mathbf{2.5 \times 10^{-6}}$	$\mathbf{2.1 \times 10^{-5}}$	$\mathbf{4.4 \times 10^{-6}}$	$\mathbf{6.0 \times 10^{-7}}$
Seasonality	0.136	0.897	0.757	0.754	0.530	0.304	0.109	0.909
Size	$\mathbf{1.5 \times 10^{-6}}$	0.513	0.941	**0.0175**	0.530	0.301	**0.022**	0.765
Skewness	0.670	0.459	0.757	0.754	0.558	0.344	0.708	0.941
Value	$\mathbf{1.5 \times 10^{-15}}$	0.897	$\mathbf{2.6 \times 10^{-14}}$	0.876	$\mathbf{3.9 \times 10^{-11}}$	0.304	$\mathbf{4.4 \times 10^{-17}}$	0.249

Notes: Underlined entries indicate significance at the 5% level using raw (unadjusted) p-values. **Bold** entries indicate Benjamini-Hochberg adjusted p-values below 5%. Inference relies on Satterthwaite approximations for the denominator degrees of freedom. All models are estimated via REML.

archical CAPM in a frequentist framework. In addition to specification 9, we estimate 11 and 12 by maximum likelihood. REML is not used, as it is inappropriate for comparing models with different fixed-effects structures. All specifications share the same random-effects structure, and the comparison is conducted only on the aggregated EM sample.

Models are compared using likelihood ratio tests (LRT) and information criteria and results are summarized in Appendix C. The LRT strongly rejects the null of common alphas or common betas across themes, indicating substantial heterogeneity in both abnormal returns and market exposures. Consistent with this result, AIC favors the most flexible specification, 9, which achieves the highest log-likelihood. In contrast, BIC favors the most parsimonious model due to its stronger penalty for model complexity. These results are reported in Tables A6 and A7.

4.2.2. Bayesian View

From the joint posterior draws, we compute marginal posterior summaries for the theme-level parameters. Figure 8 displays the posterior distributions of theme-specific alphas $\alpha_{c(i)}$ from Eq. (9) under informative global priors. We also report Bayesian 89% credible intervals, defined as the range containing 89% of posterior draws. Unlike frequentist confidence intervals, credible intervals admit a direct probabilistic interpretation (McElreath, 2018; Makowski et al., 2019).

Figure 8 shows that the posterior credible intervals for most theme-level alphas are entirely positive, with low leverage being the clear exception. Dispersion varies across themes: debt issuance, skewness, size, accruals and momentum exhibit wider posterior distributions, while value, investment and quality are comparatively more tightly concentrated.

Table A9 in Appendix D compares frequentist estimates with informative Bayesian posterior medians under informative global priors. Bayesian updating largely preserves the frequentist cross-sectional ranking of themes while shrinking extreme alphas and smoothing regional heterogeneity. Momentum and value remain the strongest and most stable sources of abnormal returns across regions, with Bayesian estimates close to or slightly above their frequentist counterparts. Profitability, profit growth, and quality are similarly reinforced, particularly in APAC and EMEA, whereas low leverage remains weak and is pulled toward zero. More marginal themes, such as size, skewness, investment, seasonality, and debt issuance, are substantially affected by shrinkage, with global information moderating extreme estimates

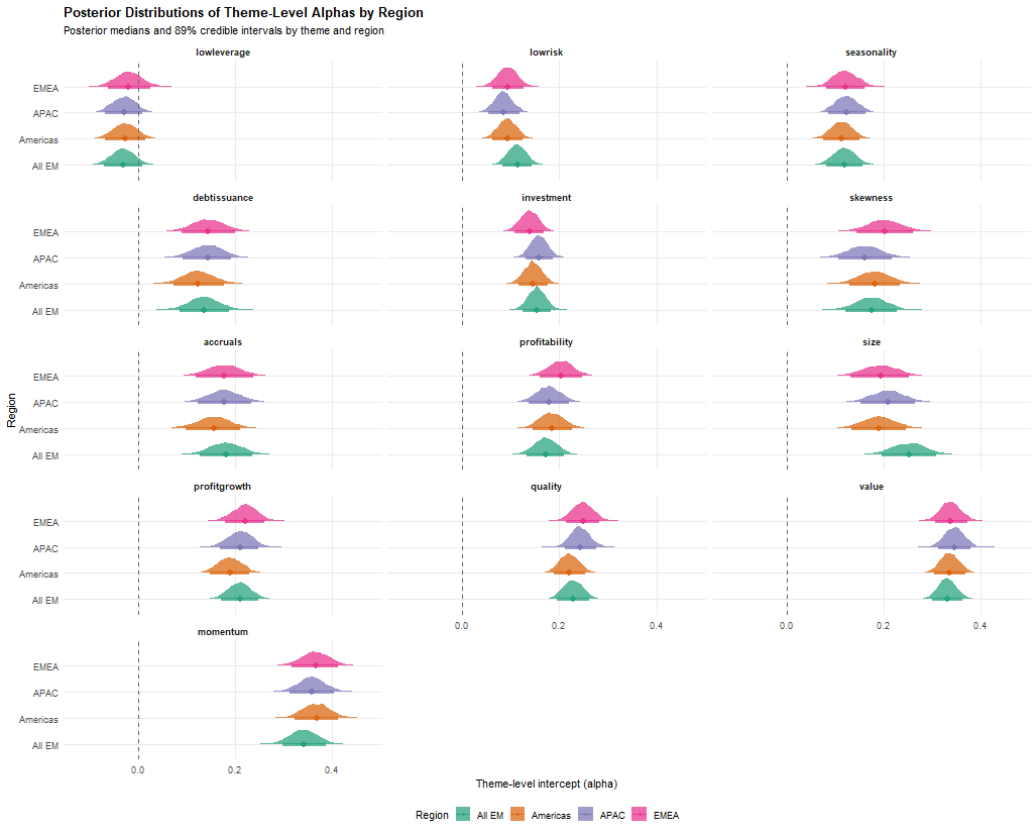

Figure 8: Posterior distribution for theme alphas across regions under informative priors derived from (Jensen et al., 2023) long-short factors aggregate global dataset . All EM and specific regions distributions are represented each with a different color and 89% credible intervals are also displayed.

and yielding more homogeneous regional patterns.

For market betas, the Bayesian hierarchy acts primarily as a shrinkage mechanism toward market neutrality. Themes with defensive exposures, most notably low risk, quality, and momentum, exhibit negative frequentist betas in some regions, which are pulled closer to zero in the posterior while preserving their relative ordering.

Figure 9 shows a markedly different pattern under conservative priors. In contrast to the informative case, most posterior theme-level alphas are shrunk toward zero. Except for profitability, quality, profit growth, momentum and value, the credible intervals of other themes extend substantially into negative territory in at least one region. Notably, these surviving themes coincide with those most extensively studied in the academic literature (Aghassi et al., 2023).

The conservative prior also induces greater dispersion and flatter posterior distributions, reflecting the relatively loose scale assumptions. Relative to the informative prior regime, this specification therefore leads to a more selective and concentrated set of themes, implying a narrower factor selection strategy.

To formally compare the evidence for the three hierarchical model specifications under the Bayesian perspective, we used the Bayes Factor approach on models defined by equations 9, 11 and 12, which share an identical random-effects structure, but differ in their fixed effects. The Bayes Factor quantifies the relative evidence provided by the data in favor of one model over another, with a value greater than 1 indicating support for the numerator model.

The analysis involved two primary tests, addressing the necessity of theme-specific fixed effects for slopes and intercepts. To save space, we report results only with the aggregated EM dataset, using the conservative prior regime. Table A8 of Appendix C reports Bayes Factors computed via bridge sampling, using the model 9 as the denominator. The strength of evidence is interpreted based on Jeffreys' scale (e.g., BF $< 1/10$ is Strong evidence against the numerator model).

Results provide decisive evidence in favor of the most flexible fixed-effects specification. In particular, the model that restricts market betas to be common across themes is overwhelmingly rejected, with a Bayes Factor of 4.53×10^{-14}, indicating essentially zero posterior support for this restriction. Similarly, the model that constrains both intercepts and market betas to be common across themes is even more strongly rejected, with a Bayes Factor of 3.23×10^{-15}.

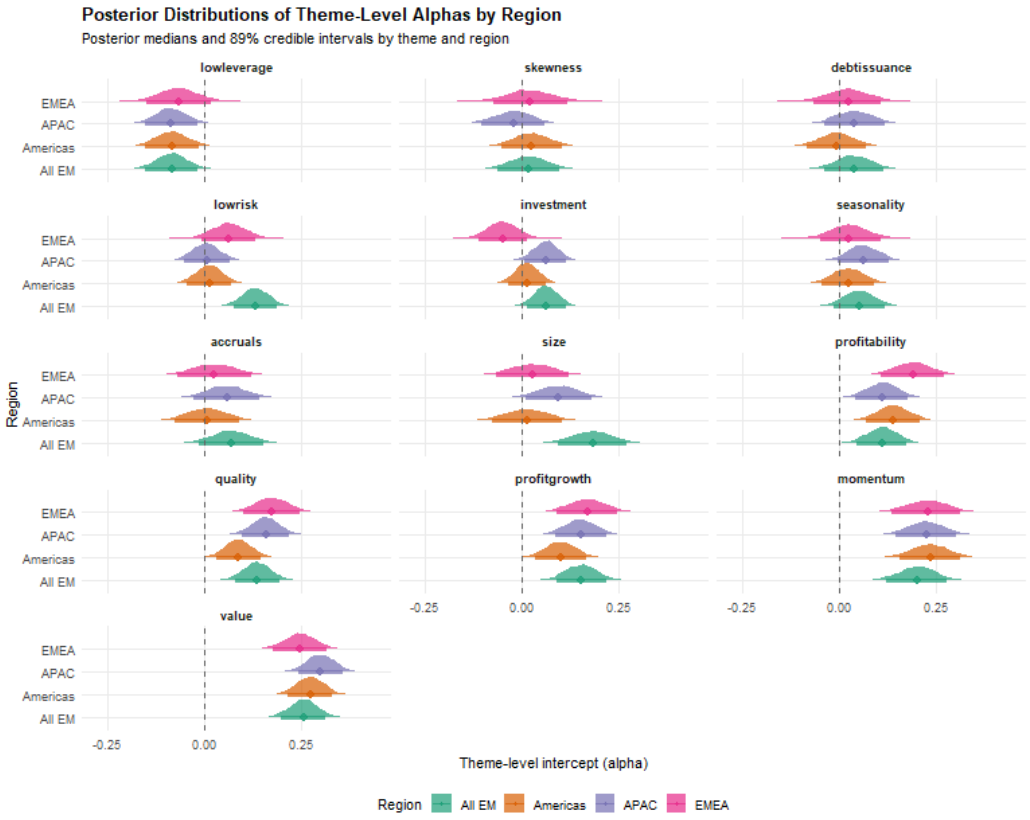

Figure 9: Posterior distribution for theme alphas across regions under conservative priors centered around 0 for $\alpha_{c(i)}$. All EM and specific regions distributions are represented each with a different color and 89% credible intervals are also displayed.

4.3. Selection Strategies Backtests

The final question is: "*can we use the hierarchical framework to improve factor selection strategies and associated out-of-sample performance and if so, what is the best way to select factors as to improve investment success?*". Using the expanding-window backtest described in Section 3.3, the following subsections summarize risk and performance metrics for emerging-market portfolios under alternative multiple-testing adjustments and weighting schemes.

4.3.1. Aggregated EM

Figure 10 illustrate the cumulative returns of all backtested strategies in Aggregated EM. To facilitate visual inspection, selection strategies were grouped by colours into 4 distinct themes: "None", "FWER", "FDR" and "Bayesian". In the Aggregated EM case, Bayesian strategies clearly outperform the remaining selection strategies, followed by FDR and FWER, with all multiple testing adjustment strategies beating the baseline no-adjustment case. Among Bayesian strategies, both prior regimes are practically equivalent.

Table 5 highlights several regularities. First, multiple-testing correction materially affects both risk and return. Although the unadjusted benchmark delivers the worst cumulative returns, it is among the least risky strategies in terms of volatility, expected shortfall and drawdowns. A natural explanation is diversification: the no-adjustment strategy selects many more factors than FWER or FDR methods (see Table 2), spreading risk across themes. However, Bayesian strategies, particularly under informative priors, are also relatively diversified yet exhibit higher risk, leading to broadly similar risk-adjusted performance across the unadjusted, FDR and Bayesian cases, with FWER methods underperforming. This pattern supports the view that FWER corrections may be overly conservative in the presence of factor multiplicity (Chen, 2021, 2022).

Among frequentist approaches, BH stands out. The BH-adjusted equal-weight portfolio achieves the best overall risk-adjusted profile, combining the highest Sharpe and Sortino ratios, the highest alpha t-statistic, the lowest downside deviation and ulcer index, and the highest Martin ratio. BH simultaneously reduces both total and downside risk while preserving strong geometric returns. In contrast, FWER methods exhibit higher volatility and drawdowns, resulting in uniformly weaker risk-adjusted metrics despite slightly higher raw returns than the unadjusted benchmark.

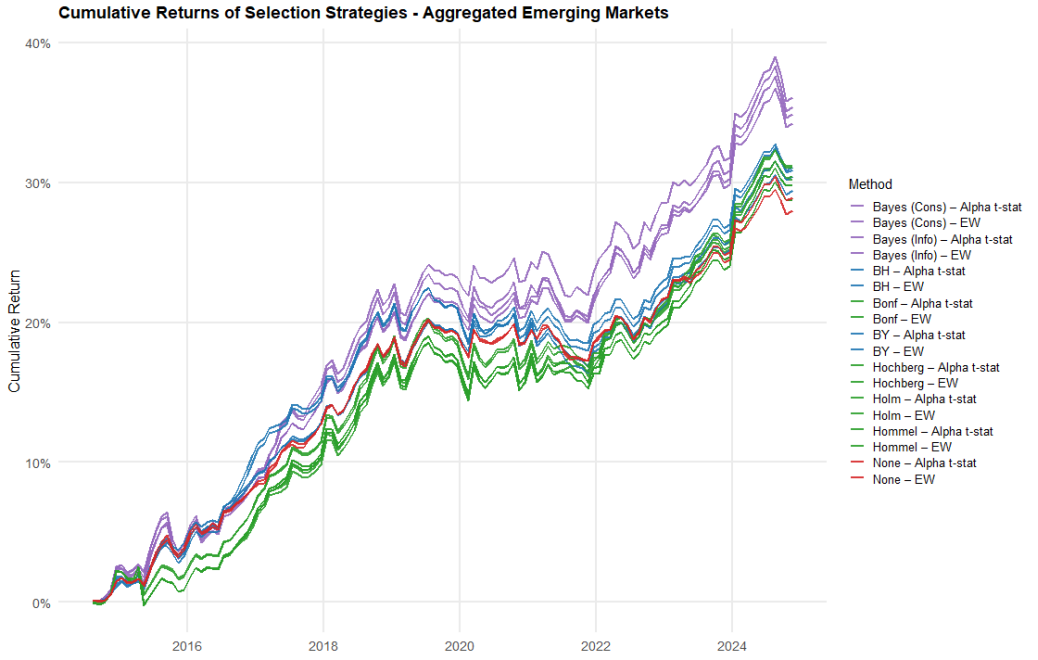

Figure 10: Cumulative returns of factor selection strategies in aggregated Emerging Markets. The figure reports the cumulative performance of portfolios constructed using different multiple-testing and Bayesian selection procedures, under equal-weight (EW) and alpha-weighted (t-statistic) schemes. Strategies are grouped by statistical approach: no adjustment (None), family-wise error rate control (FWER: Bonferroni, Holm, Hommel, Hochberg), false discovery rate control (FDR: Benjamini–Hochberg and Benjamini–Yekutieli), and Bayesian selection. Returns are compounded over time, allowing a direct comparison of long-run performance and the dynamic trade-offs between conservativeness and factor inclusion across methodologies.

Bayesian strategies display a distinct trade-off. The conservative prior delivers the highest geometric mean return and alpha, but at the cost of elevated volatility and downside risk, which weakens risk-adjusted measures. The informative global prior offers a more balanced profile, with somewhat lower raw returns but improved drawdown control, competitive Sharpe and Martin ratios, and a higher alpha t-statistic.

Finally, across both frequentist and Bayesian specifications, weighting schemes based on t_{α_i} or posterior $t^{\mathrm{B}}_{\alpha_i}$ do not systematically dominate equal weighting and sometimes worsen risk-adjusted performance. This suggests that, under mild false-discovery control, factor selection is more important than factor weighting for investment outcomes.

4.3.2. Americas

Figure 11 display the cumulative returns of all backtested strategies in Americas. Selection strategies were grouped by colours as before, but BY is not displayed since it does not select any factor given the initial buffer period, so its backtest effectively starts later than others. As before, the baseline no-adjustment case is among the selection strategies with worst out-of-sample performance, but now, Bayesian methods show the complete opposite behavior when comparing with Aggregated EM, going from best outperformance to worst outperformance. This time, we have FWER methods outperforming, followed by FDR.

Table A10 in Appendix E brings more nuances to the discussion, confirming some of the regularities documented earlier, albeit with important regional nuances. As in the All EM case, multiple-testing correction materially shapes both the risk and return profile of multifactor portfolios. However, Americas sit in a different extreme, as we have way fewer selected factor strategies on average, driving backtests' volatility considerably higher.

The unadjusted benchmark delivers low profitability and alpha but also lower risk, so do Bayesian methods, while frequentist ones show higher risk and higher returns. This time, factor diversification seems to play a more prominent role: the more permissive the selection rule is, the lower the risk and the return. While FWER methods load heavily on value and momentum with strong out-of-sample performance and 3x to 4x the alpha of the baseline case, this concentration pays a price in volatility and expected shortfall, while the selection of themes by Bayesian Informative that are completely ignored by other methods seem to help in a significant way to reduce risk, but also diminishing returns and associated alpha.

Table 5: All Emerging Markets — Risk and performance metrics by adjustment strategy and weighting scheme.

Adjustment	Scheme	Geom. Mean	Std. Dev.	Down Dev.	Sharpe	Sortino	Exp. Short	Ulcer	Martin	Alpha	Alpha T-stat
None	EW	0.1992	0.5598	0.3056	1.2460	0.6569	1.0179	0.7466	3.2366	0.1674	3.3365
	t_{α_i}	0.2051	0.5598	0.3042	1.2837	0.6795	1.0235	0.7845	3.1734	0.1757	3.4843
Bonferroni	EW	0.2134	0.6812	0.4059	1.0981	0.5314	1.4918	1.0422	2.4863	0.1967	3.1380
	t_{α_i}	0.2184	0.6829	0.4056	1.1212	0.5441	1.4943	1.0405	2.5490	0.2017	3.2100
BH	EW	0.2187	**0.5453**	**0.2877**	**1.4062**	**0.7653**	**1.0000**	**0.6159**	**4.3129**	0.1892	**3.8586**
	t_{α_i}	0.2176	0.5548	0.2957	1.3754	0.7411	1.0299	0.6552	4.0339	0.1902	3.7953
BY	EW	0.2084	0.6383	0.3677	1.1441	0.5722	1.2835	1.6777	1.5079	0.1869	3.1991
	t_{α_i}	0.2144	0.6387	0.3637	1.1767	0.5951	1.2613	1.5336	1.6976	0.1934	3.3073
Hochberg	EW	0.2041	0.6840	0.4126	1.0455	0.5004	1.5190	1.1035	2.2451	0.1880	2.9862
	t_{α_i}	0.2104	0.6853	0.4110	1.0760	0.5177	1.5151	1.0928	2.3375	0.1942	3.0794
Hommel	EW	0.2145	0.6734	0.3935	1.1164	0.5508	1.3957	1.1124	2.3413	0.1954	3.1605
	t_{α_i}	0.2193	0.6749	0.3933	1.1393	0.5634	1.4032	1.0873	2.4497	0.2001	3.2366
Holm	EW	0.2041	0.6840	0.4126	1.0455	0.5004	1.5190	1.1035	2.2451	0.1880	2.9862
	t_{α_i}	0.2104	0.6853	0.4110	1.0760	0.5177	1.5151	1.0928	2.3375	0.1942	3.0794
Bayesian (Conservative)	EW	**0.2489**	0.7293	0.4165	1.1985	0.6039	1.4039	0.8749	3.4608	0.2062	3.1633
	$t^{\mathrm{B}}_{\alpha_i}$	0.2448	0.7433	0.4271	1.1566	0.5797	1.4157	0.9674	3.0784	0.2044	3.0625
Bayesian (Global)	EW	0.2378	0.6644	0.3686	1.2565	0.6513	1.2341	0.8014	3.6086	0.2017	3.3792
	$t^{\mathrm{B}}_{\alpha_i}$	0.2417	0.6775	0.3788	1.2526	0.6442	1.2679	0.8165	3.6006	**0.2068**	3.3883

For each region, portfolios are constructed using the weighting schemes discussed in Section 3.3: equal weights (EW) or weights proportional to the frequentist t_{α_i} or posterior $t^{\mathrm{B}}_{\alpha_i}$. Selection strategies include baseline no p-value adjustment (None), Benjamini–Hochberg (BH), Benjamini–Yekutieli (BY), Bonferroni, Hommel, Holm, Bayesian with a conservative prior regime, and Bayesian with an informative global prior regime. All values are reported in monthly percentage points. Best-performing strategies for each metric are shown in **bold**.

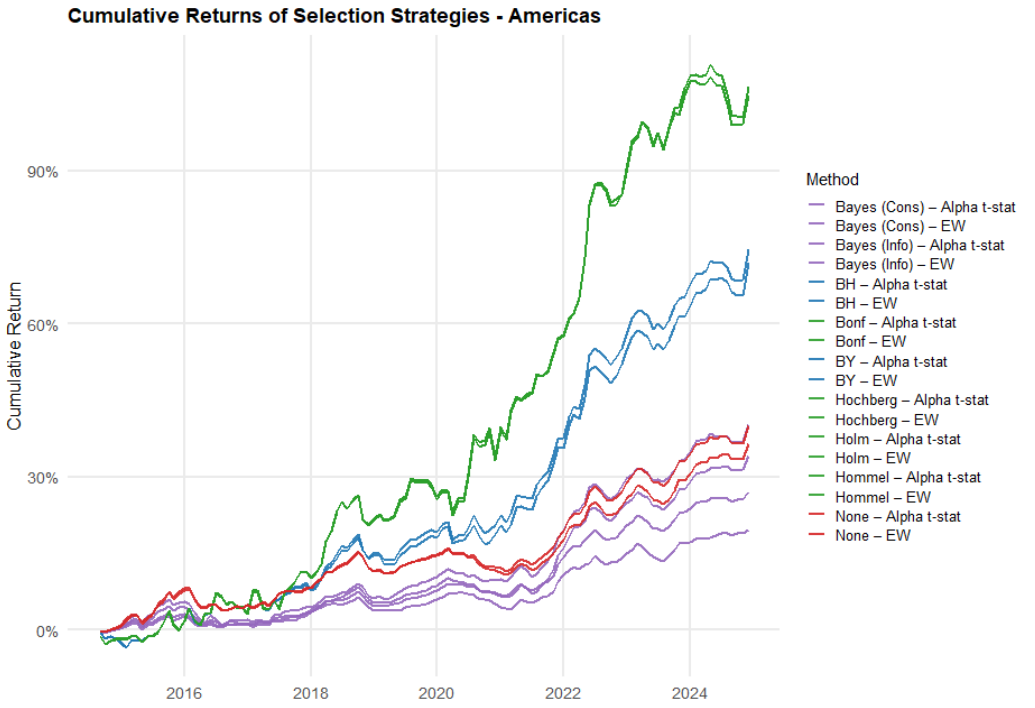

Figure 11: Cumulative returns of factor selection strategies in Americas. The figure reports the cumulative performance of portfolios constructed using different multiple-testing and Bayesian selection procedures, under equal-weight (EW) and alpha-weighted (t-statistic) schemes. Strategies are grouped by statistical approach: no adjustment (None), family-wise error rate control (FWER: Bonferroni, Holm, Hommel, Hochberg), false discovery rate control (FDR: Benjamini–Hochberg and Benjamini–Yekutieli), and Bayesian selection. Returns are compounded over time, allowing a direct comparison of long-run performance and the dynamic trade-offs between conservativeness and factor inclusion across methodologies.

All in all, BY is the selection strategy with highest alpha t-stat. Among frequentist procedures, BH again emerges as a strong compromise between signal retention and risk control, but with less dominance than in the aggregate EM results. Even so, BH displays roughly double the alpha and a higher alpha t-stat than baseline. On average, FDR methods have similar results as FWER, with BY having higher average returns and alpha than BH and FWER, but also higher risk, and BH showing smaller alpha than FWER, but also way less risk.

In terms of risk, Bayesian methods dominate, especially for the informative regime, but with a cost of average returns erosion. On average, the conservative regime display a better risk-adjusted performance than baseline, but are dominated by frequentist methods.

Finally, consistent with the ALL EM results, differences between EW and t-based weighting are second order relative to the impact of the selection rule itself.

4.3.3. APAC

Figure 12 shows backtest results in APAC, with selection strategies grouped by colours as before. As before, the baseline no-adjustment case is among the selection strategies with worst out-of-sample performance, but this time, FDR were also among the less performatic, with Bayesian and FWER showing the strongest performance.

As before, multiple-testing adjustment materially affects both profitability, alpha and drawdown behavior, and, again, unadjusted benchmark delivers weak returns and alpha, but also lower risk. Once more, despite offering higher returns and alpha, FWER procedures have higher risk due to overpruning behavior, negatively impacting risk-adjusted metrics. As a result, performance ratios and alpha t-stat remain inferior.

Offering better risk control, FDR shows improvements in risk-adjusted returns, with BY being in line with baseline. This time, Bayesian approaches stand out clearly, with the informative prior dominating the risk-adjusted landscape, achieving lowest volatility and downside risk, and also the highest performance ratios and alpha t-stat across all strategies. Conservative Bayes showed risk-adjusted metrics in line with baseline and BY. Consistent with earlier findings, weighting by frequentist or Bayesian t-statistics does not systematically outperform equal weighting.

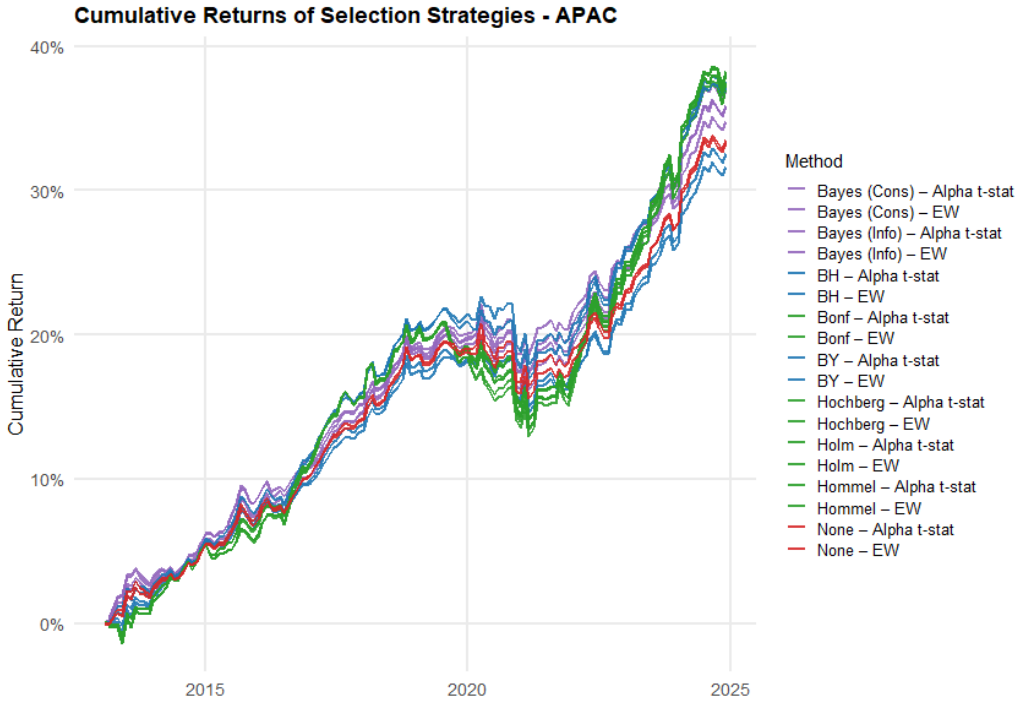

Figure 12: Cumulative returns of factor selection strategies in APAC. The figure reports the cumulative performance of portfolios constructed using different multiple-testing and Bayesian selection procedures, under equal-weight (EW) and alpha-weighted (t-statistic) schemes. Strategies are grouped by statistical approach: no adjustment (None), family-wise error rate control (FWER: Bonferroni, Holm, Hommel, Hochberg), false discovery rate control (FDR: Benjamini–Hochberg and Benjamini–Yekutieli), and Bayesian selection. Returns are compounded over time, allowing a direct comparison of long-run performance and the dynamic trade-offs between conservativeness and factor inclusion across methodologies.

4.3.4. EMEA

Figure 13 shows cumulative returns of all backtested strategies in EMEA. This time, frequentist adjustment methods clearly outperform baseline and Bayesian strategies.

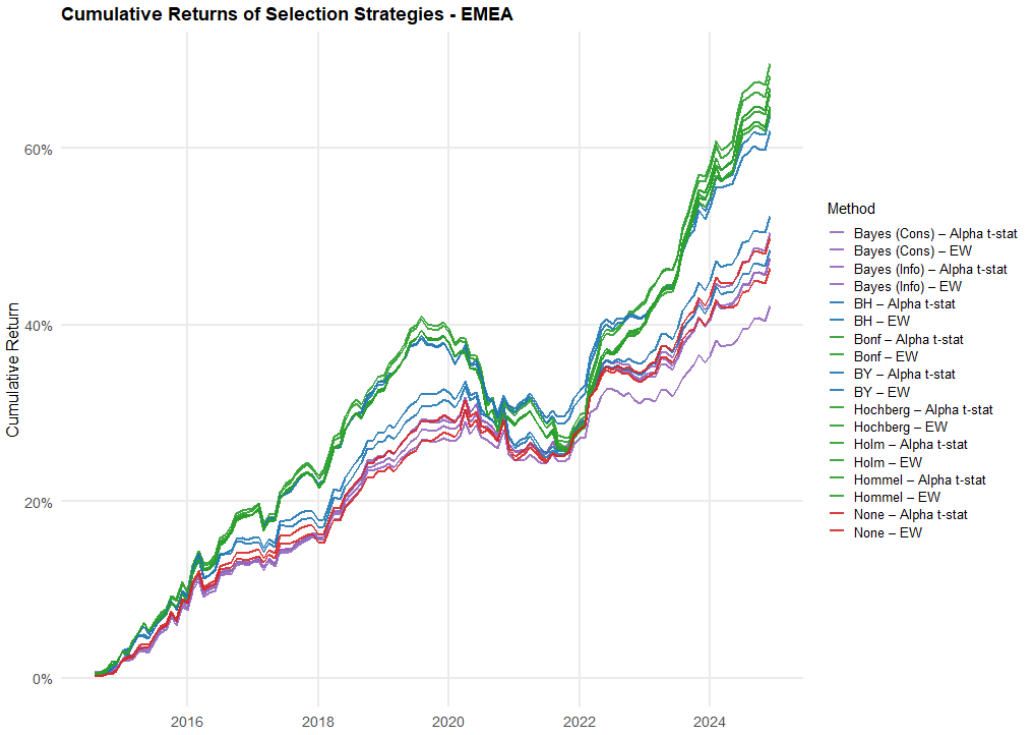

Figure 13: Cumulative returns of factor selection strategies in EMEA. The figure reports the cumulative performance of portfolios constructed using different multiple-testing and Bayesian selection procedures, under equal-weight (EW) and alpha-weighted (t-statistic) schemes. Strategies are grouped by statistical approach: no adjustment (None), family-wise error rate control (FWER: Bonferroni, Holm, Hommel, Hochberg), false discovery rate control (FDR: Benjamini–Hochberg and Benjamini–Yekutieli), and Bayesian selection. Returns are compounded over time, allowing a direct comparison of long-run performance and the dynamic trade-offs between conservativeness and factor inclusion across methodologies.

EMEA results complete the regional picture and reinforce the main cross-market conclusions: the unadjusted benchmark delivers the weakest performance and alpha, while multiple-testing adjustment generally increases risk, but leads to better risk-adjusted returns and alpha t-statistics.

FWER strategies again exhibit the highest risk. In EMEA, however, the rise in volatility and downside risk is not sufficient to offset their higher returns in Sharpe and Sortino terms, although the Martin ratio deteriorates due to higher ulcer indices. Alpha levels and alpha t-statistics are also materially stronger under FWER corrections.

On average, FDR improves clearly over the unadjusted benchmark in terms of risk-adjusted performance and alpha, though its dominance is less pronounced than in the All Emerging sample. BY generates best alpha t-statistic, but BH is actually wrose than no adjustment case.

Bayesian approaches again display a distinct profile, with informative prior delivering the strongest risk-adjusted performance and higher alpha t-statistic than baseline.

4.3.5. Consolidated findings

Taken together, the results across Aggregated EM, Americas, APAC, and EMEA provide a clear and coherent answer to the central question of this subsection: hierarchical factor selection frameworks materially impacts out-of-sample performance, but the optimal selection rule is region-dependent and also depends on individual investor risk preferences.

Figure 14 plots CAPM alpha (y-axis) against specific risk (x-axis) for each out-of-sample backtest portfolio and Table 6 averages out-of-sample metrics across regions and weighting schemes.

A first conclusion is that the baseline no-adjustment strategy out-of-sample alpha can usually be improved by applying adjustments. However, this comes with a clear cost of higher risk, although risk-adjusted ratios are usually higher for the adjustment case, especially in Americas and EMEA. The trade-off was more extreme at the Americas regions, with the choice of adjustment method providing material shift across risk-return spectrum and number of selected factors.

Regarding the choice of multiple adjustment method, FWER controls consistently yield strategies with highest alphas. However, most of the time, this makes FWER adjustments have worse risk-adjusted return metrics than benchmark and other strategies, but improved alpha t-stat compared to the former. On the other hand, FDR produce more nuanced outcomes and, on average, outperform all others in risk-adjusted returns and alpha t-stats. As for Bayesian adjustments, while the conservative priors shows, on average, similar out-of-sample behavior as baseline, the informative prior regime can offer risk-adjusted improvements, usually through risk reduction, as this

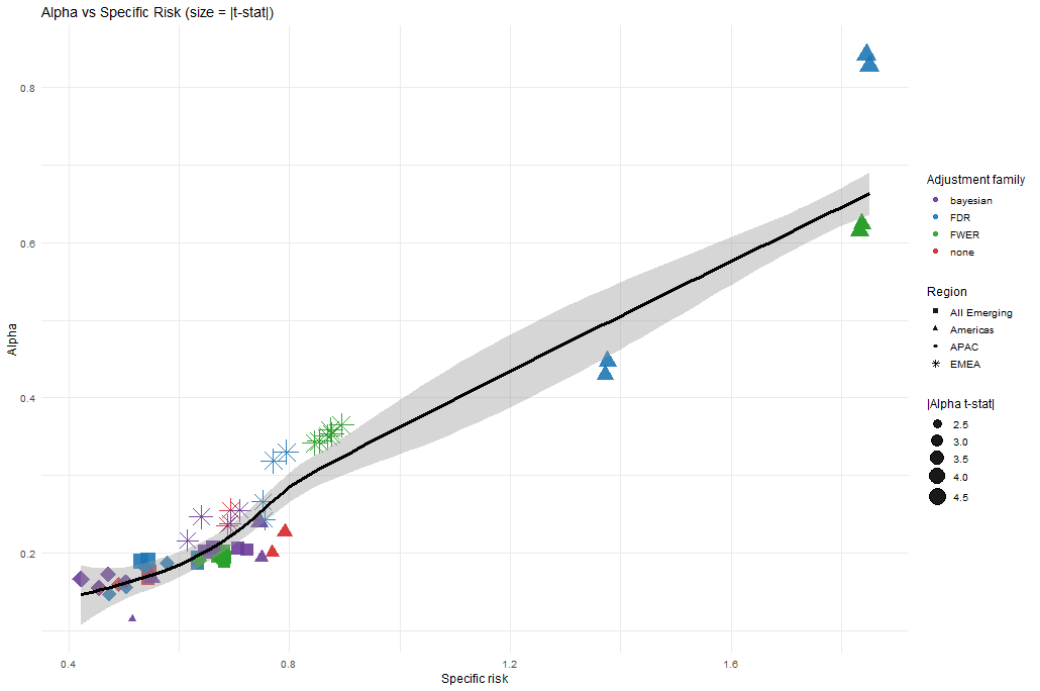

Figure 14: **Alpha vs specific-risk across selection rules and regions.** Scatter plot of out-of-sample strategy CAPM alphas versus CAPM specific risk. Each point corresponds to one backtested selection strategy. Marker size is proportional to the absolute alpha t-statistic, marker shape identifies the region (All Emerging, Americas, APAC, EMEA), and marker color denotes the selection/adjustment family: none (no adjustment), FDR control (BH/BY), FWER control (Bonferroni/Holm/Hochberg/Hommel), and Bayesian rules. The black line is a nonlinear GAM fit of $\hat{\alpha}$ on specific risk with a 95% confidence band.

method consistently shows the lowest risk overall.

This picture leads to a second conclusion: while adjusting p-values is usually a good idea, the exact adjustment depends on both the region, without an obvious best method that always outperform, and on individual utility function. When choosing high alpha and returns, FWER is the way to go, while, for much risk averse investors, Bayesian shrinkage with global priors is a better move and, finally, FDR methods can offer the best risk-adjusted metrics.

Finally, our last conclusion is that weighting schemes play a secondary role. Across all regions and methodologies, weighting portfolios by frequentist or Bayesian t-statistics does not systematically outperform equal weighting. Differences between EW and t-based schemes are small relative to the impact of the selection rule itself, reinforcing the conclusion that getting the set of signals right matters more than fine-tuning their weights.

Table 6: Consolidated performance across factor-selection paradigms

Metric	None	FWER	FDR	Bayes Cons.	Bayes Info.
Geom. Mean	0.245	**0.357**	0.341	0.258	0.231
Std. Dev.	0.661	1.020	0.900	0.705	**0.603**
Down Dev.	0.367	0.552	0.473	0.388	**0.329**
Ann. Sharpe	1.31	1.25	**1.35**	1.29	1.34
Sortino	0.678	0.651	**0.721**	0.675	0.707
Exp. Short	1.36	1.91	1.68	1.42	**1.19**
Ulcer	1.26	1.99	1.47	1.29	**0.98**
Martin	2.54	2.35	2.97	2.58	**3.02**
Alpha	0.197	**0.340**	0.321	0.210	0.187
Alpha t-stat	3.54	3.66	**3.77**	3.50	3.68

Values are averages across all regions (All EM, APAC, Americas, EMEA) and weighting schemes. Bold values indicate best performance per metric (return, risk, or risk-adjusted).

5. Conclusion

This paper reexamines the factor zoo through the lens of multiple testing and hierarchical modeling, motivated by concerns over p-hacking, false discoveries, and post-publication decay. Using Emerging Markets as a demanding and economically relevant out-of-sample laboratory, given their relative

insulation from factor crowding and data-mining, combined with harsher inferential conditions, we study factor alphas existence and magnitude, how they should be modeled statistically, and whether disciplined selection frameworks can be translated into superior out-of-sample portfolio performance.

Our first set of results addresses the existence and magnitude of factor alphas. Using no-pooled and hierarhical CAPM regressions, we document that a small subset of themes, most notably value, profitability, quality, profit growth, and momentum, concentrates the most persistent and economically meaningful alphas across regions. At the same time, several commonly cited themes display weak abnormal returns, and, notably, factors in the low leverage theme presented frequent significant negative alphas.

Region-wise, the Americas showed the smallest number of significant factors and themes, with a concentration on value and momentum, and, on the other hand APAC and the aggregated dataset displayed the highest, favoring themes such as size, accruals, low risk, debt issuance, seasonality and investment. As for choice of inference method, FWER methods naturally provided the most conservative selection rule, with significant alphas more concentrated under more prominent themes. At the same time, evaluating alphas under Bayesian inference with informative global priors resulted in more support for factors in other themes, except for low leverage, having a larger impact in the Americas.

The absence of robust premia for some commonly cited themes admits at least three non-mutually exclusive explanations. First, these signals may reflect data mining or p-hacking in developed-market samples, failing to survive out-of-sample tests in EM. Second, EM may differ fundamentally in market structure, investor composition, regulation, or firm characteristics, such that certain mechanisms underlying factor premia in developed markets are attenuated or absent. Third, some themes may be subject to long cycles or regime dependence and could re-emerge in EM with larger samples. Distinguishing among these channels is a recommendation for future studies.

Our second contribution concerns the role of hierarchical structure in factor inference. Modeling factors as nested within broader economic themes result in a more realistic and parsimonious framework than a no-pooled independent representation. In special, Bayesian inference, which is receiving a growing amount of support as a possible solution for the loss in credibility of some findings in given fields, including asset pricing, provides an increased advantage in EM, because of their more volatile, data-scarce and shallower markets. As for specification, there is evidence of varying fixed effects for

both alpha and market factor exposures across themes, with similar conclusions in terms of which themes dominate the cross-sectional alpha generation in EM. In addition, we document significant negative market factor exposure for themes such as low risk, quality, momentum, and investment and neutrality for the rest. Future research may extend our specification with dynamic parameters and priors. Finally, we point out that hierarchical lens provide support for an additional number of themes when compared to the no-pooled perspective.

The third and most economically relevant set of findings links statistical inference to out-of-sample portfolio performance. Across all regions and, especially in Americas and EMEA, ignoring multiple testing leads to inferior out-of-sample risk-adjusted outcomes. Although unadjusted strategies often appear diversified and low risk, they deliver weak returns and alpha, and applying multiple testing adjustments in selecting factors can materially change both risk and return, with evidence of improvements in risk-adjusted ratios and alpha t-stats.

However, the choice of multiple control method itself depends more on region and risk preferences. At one extreme, family-wise error rate (FWER) controls generate the highest raw returns and alphas, but entail high risk and aggressively concentrating portfolios, resulting in often worse risk-adjusted out-of-sample performance. On the other hand, Bayesian methods, especially with informative priors, can offer more diversification and less risk, especially in APAC and EMEA, improving risk-adjusted results. Finally, false discovery rate (FDR) control offered the best ratios and alpha t-stats overall. We conclude that there is not a "one-size-fits-all" approach. Finally, we find little systematic advantage to weighting schemes based on frequentist or Bayesian t-statistics relative to equal weighting. Therefore, differences in sizing are second-order compared to controlling false discoveries.

Ultimately, we point out, albeit more concentrated in fewer themes, factors do exhibit statistical significant alphas in EM, with some themes showing robust performance regardless of multiple correction or specification choice. A hierarchical view can serve as a more coherent and parsimonious framework to the factor zoo and there is evidence that both abnormal returns and market exposures vary across themes. In fact, diversifying across factors lead to improvements in risk reduction. Finally, disciplined inference is not a constraint on asset pricing, but a prerequisite for extracting its true economic value and shaping portfolio risk-return to individual preferences.

References

Aghassi, M. et al. (2023), 'Fact, fiction and factor investing', *The Journal of Portfolio Management* **49**, 1–38.

Asness, C. (2011), 'Momentum in japan: The exception that proves the rule', *The Journal of Portfolio Management* **37**, 67–75.

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

Bailey, D. H., Borwein, J., López de Prado, M. and Zhu, Q. J. (2014), 'Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample performance', *Notices of the American Mathematical Society* **61**(5), 458–471. Originally posted August 2013; last revised April 2014.

Bailey, D. H. and López de Prado, M. (2014), 'The deflated sharpe ratio: Correcting for selection bias, backtest overfitting, and non-normality', *The Journal of Portfolio Management* **40**(5), 94–107.

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

Bekaert, G. and Harvey, C. R. (1997), 'Emerging equity market volatility', *Journal of Financial Economics* **43**, 29–77.

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

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

Bergmeir, C., Hyndman, R. and Koo, B. (2018), 'A note on the validity of cross-validation for evaluating autoregressive time series prediction', *Computational Statistics & Data Analysis* **120**, 70–83.

Berkson, J. (1942), 'Tests of significance considered as evidence', *Journal of the American Statistical Association* **37**, 325–335.

Cakizi, N., Fabozzi, F. and Tan, S. (2013), 'Size, value, and momentum in emerging market stock returns', *Emerging Markets Review* **16**, 46–65.

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

Chen, A. (2022), 'Do t-statistic hurdles need to be raised', *arXiv.org* **2204.10275**.

Chen, A. and Zimmermann, T. (2020), 'Publication bias and the cross-section of stock returns', *The Review of Asset Pricing Studies* **10**, 249–289.

Chordia, T., Goyal, A. and Saretto, A. (2020), 'Anomalies and false rejections', *The Review of Financial Studies* **33**, 2134–2179.

Cochrane, J. (2011), 'Discount rates', *NBER Working Papers* **16972**.

Cohen, J. (1994), 'The earth is round ($p < .05$)', *American Psychologist* **49**, 997–1003.

De Prado, M. (2023), 'Causal factor investing: Can factor investing become scientific?', *Available at SSRN: https://ssrn.com/abstract=4205613* .

Fama, E. and French, K. (2017), 'International tests of a five-factor asset pricing model', *Journal of Financial Economics* **123**, 441–463.

Fanelli, D. (2010), '"positive" results increase down the hierarchy of the sciences', *PLoS ONE* **5**, e10068.

Gelman, A. (2015), 'The connection between varying treatment effects and the crisis of unreplicable research: A bayesian perspective', *Journal of Management* **41**(2), 632–643.

Goyal, A. (2012), 'Empirical cross-sectional asset pricing: A survey', *Financial Markets and Portfolio Management* **26**, 3–38.

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

Gu, S., Kelly, B. and Dacheng, X. (2020), 'Empirical asset pricing via machine learning', *The Review of Financial Studies* **33**, 2223–2273.

Guimarães, P. R. and Kimura, H. (2025), '(re)emergence of the factor zoo debate: A bibliometric review', *International Journal of Emerging Markets* **20**(11), 4597–4621. Literature Review, September 25, 2024.

Hanauer, M. and Lauterbach, J. (2019), 'The cross-section of emerging market stock returns', *Emerging Markets Review* **38**, 265–286.

Hanauer, M. X. and Linhart, M. (2015), 'Size, value, and momentum in emerging market stock returns: Integrated or segmented pricing?', *Asia-Pacific Journal of Financial Studies* **44**(2), 175–214.

Harvey, C. (1995), 'Predictable risk and returns in emerging markets', *The Review of Financial Studies* **8**, 773–816.

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

Harvey, C. (2019), 'Editorial: Replication in financial economics', *Critical Finance Review* **8**, 1–9.

Harvey, C. and Liu, Y. (2019), 'A census of the factor zoo', *Available at SSRN: https://ssrn.com/abstract=3341728* .

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

Harvey, C. R. (1994), Predictable risk and returns in emerging markets, NBER Working Paper 4621, National Bureau of Economic Research.

Hau, H. (2011), 'Global versus local asset pricing: A new test of market integration', *The Review of Financial Studies* **24**, 3891–3940.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T. and Jennions, M. D. (2015), 'The extent and consequences of p-hacking in science', *PLoS Biology* **13**(3), e1002106.

Hochberg, Y. (1988), 'A sharper bonferroni procedure for multiple tests of significance', *Biometrika* **75**(4), 800–802.

Hochberg, Y. and Tamhane, A. C. (1987), *Multiple Comparison Procedures*, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, John Wiley & Sons, New York.

Holm, S. (1979), 'A simple sequentially rejective multiple test procedure', *Scandinavian Journal of Statistics* **6**(2), 65–70.

Hommel, G. (1988), 'A stagewise rejective multiple test procedure based on a modified bonferroni test', *Biometrika* **75**, 383–386.

Ilmanen, A. et al. (2021), 'How do factor premia vary over time? a century of evidence', *Journal of Investment Management* **19**, 15–57.

Ioannidis, J. P. A. (2018), 'What have we (not) learnt from millions of scientific papers with p values?', *The American Statistician* .

Jacobs, H. and Müller, S. (2020), 'Anomalies across the globe: Once public, no longer existent?', *Journal of Financial Economics* **135**, 213–230.

Jensen, T., Kelly, B. and Pedersen, L. (2023), 'Is there a replication crisis in finance?', *The Journal of Finance* **78**, 2465–2518.

Johnson, A. A., Ott, M. Q. and Dogucu, M. (2022), *Bayes Rules!: An Introduction to Applied Bayesian Modeling*, 1 edn, Chapman and Hall/CRC, Boca Raton.

Karolyi, A. (2016), 'Home bias, an academic puzzle', *Review of Finance* **20**, 2049–2078.

Kuznetsova, A., Brockhoff, P. B. and Christensen, R. H. B. (2017), 'lmerTest package: Tests in linear mixed effects models', *Journal of Statistical Software* **82**(13), 1–26.

Maiti, M. (2019), 'A critical review on evolution of risk factors and factor models', *Journal of Economic Surveys* **34**, 175–184.

Makowski, D., Ben-Shachar, M. S. and Lüdecke, D. (2019), 'bayestestr: Describing effects and their uncertainty, existence and significance within the bayesian framework', *Journal of Open Source Software* **4**(40), 1541.
URL: *https://doi.org/10.21105/joss.01541*

McElreath, R. (2018), *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*, Chapman & Hall/CRC.

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

Meehl, P. E. (1978), 'Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology', *Journal of Consulting and Clinical Psychology* **46**, 806–834. Based on a lecture delivered at the American Psychological Association meeting, Washington, D.C., September 1976. Research supported by NIMH Grant MH 24224.

Mody, A. (2004), 'What is an emerging market?', *IMF Working Papers No. 2004/177* .

Morck, R., Yeung, B. and Yu, W. (2000), 'The information content of stock markets: Why do emerging markets have synchronous stock price movements?', *Journal of Financial Economics* **58**(1–2), 215–260.

MSCI (2023), 'MSCI market classification framework'. Accessed = 2024-05-19.
URL: *https://www.msci.com/documents/1296102/e7613c4f-f16c-b039-f69a-fbdaafed579f*

MSCI Inc. (2025), Msci market classification framework, Index framework, MSCI. Accessed: 25 November 2025.

Rapach, D. E., Strauss, J. K. and Zhou, G. (2013), 'International stock return predictability: What is the role of the united states?', *The Journal of Finance* **68**(4), 1633–1662.

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

Rozeboom, W. W. (1960), 'The fallacy of the null-hypothesis significance test', *Psychological Bulletin* **57**(5), 416–428.

Schnaubelt, M. (2019), A comparison of machine learning model validation schemes for non-stationary time series data, Technical report, FAU Discussion Papers in Economics.

Simes, R. J. (1986), 'An improved bonferroni procedure for multiple tests of significance', *Biometrika* **73**, 751–754.

Tashman, L. (2000), 'Out-of-sample tests of forecasting accuracy: An analysis and review', *International Journal of Forecasting* **16**, 437–450.

van Zwet, E. and Cator, E. (2021), 'The significance filter, the winner's curse and the need to shrink', *Statistica Neerlandica* .

Zaremba, A. and Czapkiewicz, A. (2017), 'Digesting anomalies in emerging european markets: A comparison of factor pricing models', *Emerging Markets Review* **31**, 1–15.

Appendix

Appendix A. Additional Frequentist Results

Appendix A.1. Theme-Level Selection Rates by Region

This section reports theme-level proportions of selected factors across regions and multiple-testing adjustments, complementing the aggregate results discussed in the main text.

Table A1: Proportion of significant factors by theme in the JKP emerging markets cluster

Theme	None	Bonferroni	Holm	Hochberg	Hommel	BH	BY
Accruals	0.600	0.400	0.600	0.600	0.600	0.600	0.600
Debt issuance	0.429	0.143	0.143	0.143	0.143	0.286	0.143
Investment	0.455	0.046	0.046	0.046	0.091	0.318	0.136
Low leverage	0.200	0.000	0.000	0.000	0.000	0.100	0.000
Low risk	0.294	0.059	0.059	0.059	0.059	0.235	0.235
Momentum	0.750	0.125	0.125	0.125	0.125	0.625	0.250
Profitability	0.545	0.182	0.182	0.182	0.182	0.364	0.182
Profit growth	0.727	0.455	0.455	0.455	0.455	0.545	0.455
Quality	0.562	0.000	0.063	0.063	0.063	0.375	0.125
Seasonality	0.364	0.182	0.182	0.182	0.182	0.273	0.182
Size	0.800	0.200	0.200	0.200	0.200	0.800	0.600
Skewness	0.167	0.000	0.000	0.000	0.000	0.000	0.000
Value	0.778	0.333	0.333	0.333	0.333	0.667	0.444

This table reports, for the JKP emerging markets cluster, the share of characteristics within each economic theme that are significant at the 5% level under alternative multiple-testing adjustments. "None" refers to unadjusted p-values; Bonferroni, Holm, Hochberg and Hommel control the family-wise error rate (FWER); BH and BY denote the Benjamini-Hochberg and Benjamini-Yekutieli false discovery rate (FDR) procedures.

Appendix B. Additional Bayesian Results

Appendix B.1. Theme-Level Posterior Probabilities

This section reports posterior probabilities of positive alphas by theme under alternative prior regimes.

Table A2: Proportion of significant factors by theme in the JKP APAC emerging markets cluster

Theme	None	Bonferroni	Holm	Hochberg	Hommel	BH	BY
Accruals	0.600	0.000	0.000	0.000	0.400	0.600	0.400
Debt issuance	0.571	0.143	0.143	0.143	0.143	0.571	0.286
Investment	0.409	0.091	0.091	0.091	0.091	0.409	0.136
Low leverage	0.091	0.000	0.000	0.000	0.000	0.091	0.000
Low risk	0.059	0.059	0.059	0.059	0.059	0.059	0.059
Momentum	1.000	0.125	0.125	0.125	0.125	0.875	0.375
Profitability	0.636	0.182	0.182	0.182	0.182	0.455	0.182
Profit growth	0.727	0.364	0.364	0.364	0.364	0.727	0.545
Quality	0.706	0.059	0.118	0.118	0.118	0.588	0.294
Seasonality	0.455	0.000	0.091	0.091	0.091	0.364	0.182
Size	0.800	0.000	0.000	0.000	0.000	0.400	0.000
Skewness	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Value	0.889	0.444	0.500	0.500	0.500	0.889	0.667

Notes: This table reports, for the JKP APAC emerging markets cluster, the proportion of characteristics within each economic theme that are classified as significant at the 5% level under different multiple-testing adjustment procedures. "None" denotes unadjusted p-values; Bonferroni, Holm, Hochberg and Hommel control the family-wise error rate (FWER); BH and BY correspond to the Benjamini-Hochberg and Benjamini-Yekutieli false discovery rate (FDR) procedures, respectively.

Table A3: Proportion of significant factors by theme in the JKP Americas emerging markets cluster

Theme	None	Bonferroni	Holm	Hochberg	Hommel	BH	BY
Accruals	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Debt issuance	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Investment	0.046	0.000	0.000	0.000	0.000	0.000	0.000
Low leverage	0.111	0.000	0.000	0.000	0.000	0.000	0.000
Low risk	0.059	0.000	0.000	0.000	0.000	0.000	0.000
Momentum	0.875	0.250	0.250	0.250	0.250	0.375	0.250
Profitability	0.364	0.000	0.000	0.000	0.000	0.182	0.000
Profit growth	0.250	0.000	0.000	0.000	0.000	0.083	0.000
Quality	0.235	0.000	0.000	0.000	0.000	0.059	0.000
Seasonality	0.091	0.000	0.000	0.000	0.000	0.000	0.000
Size	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Skewness	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Value	0.722	0.222	0.222	0.222	0.222	0.389	0.222

Notes: This table reports, for the JKP Americas emerging markets cluster, the proportion of characteristics within each economic theme that are classified as significant at the 5% level under different multiple-testing adjustment procedures. "None" denotes unadjusted p-values; Bonferroni, Holm, Hochberg and Hommel control the family-wise error rate (FWER); BH and BY correspond to the Benjamini-Hochberg and Benjamini-Yekutieli false discovery rate (FDR) procedures, respectively.

Table A4: Proportion of significant factors by theme in the JKP EMEA emerging markets cluster

Theme	None	Bonferroni	Holm	Hochberg	Hommel	BH	BY
Accruals	0.200	0.000	0.000	0.000	0.000	0.200	0.000
Debt issuance	0.286	0.000	0.000	0.000	0.000	0.000	0.000
Investment	0.091	0.000	0.000	0.000	0.000	0.000	0.000
Low leverage	0.100	0.100	0.100	0.100	0.100	0.100	0.100
Low risk	0.250	0.000	0.000	0.000	0.000	0.250	0.000
Momentum	1.000	0.000	0.000	0.000	0.000	1.000	0.375
Profitability	0.727	0.273	0.273	0.273	0.273	0.545	0.273
Profit growth	0.583	0.250	0.250	0.250	0.250	0.417	0.250
Quality	0.688	0.063	0.063	0.063	0.063	0.375	0.312
Seasonality	0.091	0.000	0.000	0.000	0.000	0.000	0.000
Size	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Skewness	0.200	0.000	0.000	0.000	0.000	0.000	0.000
Value	0.667	0.278	0.278	0.278	0.278	0.500	0.333

Notes: This table reports, for the JKP EMEA emerging markets cluster, the proportion of characteristics within each economic theme that are classified as significant at the 5% level under different multiple-testing adjustment procedures. "None" denotes unadjusted p-values; Bonferroni, Holm, Hochberg and Hommel control the family-wise error rate (FWER); BH and BY correspond to the Benjamini–Hochberg and Benjamini–Yekutieli false discovery rate (FDR) procedures, respectively.

Table A5: Posterior evidence of positive alphas by theme

	All emerging		Americas		APAC		EMEA	
Theme	Conservative	Informative	Conservative	Informative	Conservative	Informative	Conservative	Informative
accruals	0.600	**0.600**	0.000	**0.400**	0.600	**0.600**	0.200	**0.200**
debtissuance	0.143	**0.143**	0.000	**0.000**	0.286	**0.429**	0.000	**0.286**
investment	**0.227**	**0.455**	0.000	**0.273**	**0.409**	**0.409**	0.000	**0.091**
lowleverage	0.000	0.000	0.000	0.000	0.091	0.091	0.100	0.100
lowrisk	**0.706**	**0.706**	0.000	**0.059**	0.059	**0.059**	0.250	**0.250**
momentum	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
profitability	**0.545**	**0.545**	**0.909**	**0.909**	**0.636**	**0.636**	**0.727**	**0.818**
profitgrowth	**0.636**	**0.818**	**0.417**	**0.750**	**0.818**	**0.818**	**0.667**	**0.750**
quality	**0.750**	**0.938**	**0.176**	**0.941**	**0.824**	**0.882**	**0.750**	**0.812**
seasonality	0.273	**0.273**	0.000	**0.091**	0.273	**0.455**	0.000	**0.091**
size	**1.000**	**1.000**	0.000	**0.500**	**0.800**	**0.800**	0.000	**0.200**
skewness	0.167	**0.333**	0.000	**0.667**	0.000	**0.000**	0.000	**0.200**
value	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**0.833**	**0.889**

Entries are the proportion of individual factor alphas with $p_d \geq 0.95$ within each theme and region. Bold values indicate that the corresponding theme-level alpha also has $p_d \geq 0.95$.

Appendix C. Model Comparison and Specification Tests

Appendix C.1. Frequentist Specifications Comparison

This section reports likelihood-ratio tests and information criteria used to compare alternative frequentist hierarchical specifications.

Table A6: Likelihood ratio tests for hierarchical CAPM specifications

Comparison	k_1	k_2	χ^2	df	p-value
α_i, β_i (eq. 12) vs. $\alpha_{c(i)}, \beta_{c(i)}$ (eq. 9)	6	30	175.29	24	$< 2.2 \times 10^{-16}$
$\alpha_{c(i)}, \beta_i$ (eq. 11) vs. $\alpha_{c(i)}, \beta_{c(i)}$ (eq. 9)	18	30	99.24	12	7.85×10^{-16}

k_1 and k_2 denote the number of estimated parameters in the restricted and unrestricted models, respectively.

Table A7: Model comparison via likelihood and information criteria

Model	Log-likelihood	Deviance	AIC	BIC
α_i, β_i (eq. 9)	−66184.23	132368.5	132380.5	132431.2
$\alpha_{c(i)}$, β_i (eq. 11)	−66146.15	132292.3	132328.3	132480.6
$\alpha_{c(i)}$, $\beta_{c(i)}$ (eq. 12)	−66096.53	132193.1	132253.1	132506.9

Lower values of AIC and BIC indicate better model fit after penalizing for model complexity. Bold values denote the minimum criterion.

Appendix C.2. Bayesian Specifications Comparison

This section displays Bayes Factor comparisons for alternative hierarchical specifications.

Table A8: Bayes Factor Comparison for Fixed Effects Structure

Num.	Denom.	Bayes Factor	Evidence for Denom.
$\alpha_{c(i)}$, β_i (eq. 11)	$\alpha_{c(i)}$, $\beta_{c(i)}$ (eq. 9)	4.53×10^{-14}	Decisive
α_i, β_i (eq. 12)	$\alpha_{c(i)}$, $\beta_{c(i)}$ (eq. 9)	3.23×10^{-15}	Decisive

Table A9: Frequentist and Bayesian theme-level CAPM alphas and betas by region

Theme	All EM				APAC				Americas				EMEA			
	$\alpha^F_{c(i)}$	$\alpha^B_{c(i)}$	$\beta^F_{c(i)}$	$\beta^B_{c(i)}$	$\alpha^F_{c(i)}$	$\alpha^B_{c(i)}$	$\beta^F_{c(i)}$	$\beta^B_{c(i)}$	$\alpha^F_{c(i)}$	$\alpha^B_{c(i)}$	$\beta^F_{c(i)}$	$\beta^B_{c(i)}$	$\alpha^F_{c(i)}$	$\alpha^B_{c(i)}$	$\beta^F_{c(i)}$	$\beta^B_{c(i)}$
accruals	0.138	0.182	-0.003	-0.009	0.119	0.178	-0.001	-0.009	0.012	0.157	-0.005	-0.009	0.068	0.178	-0.026	-0.011
debt issuance	0.063	0.136	-0.029	-0.007	0.070	0.144	-0.040	-0.007	-0.011	0.124	-0.027	-0.006	0.047	0.145	-0.029	-0.007
investment	0.078	0.154	-0.040	-0.017	0.078	0.158	-0.053	-0.017	0.017	0.146	-0.044	-0.016	-0.066	0.138	-0.020	-0.016
low leverage	-0.123	-0.030	-0.028	0.005	-0.127	-0.028	-0.031	0.006	-0.128	-0.026	-0.001	0.008	-0.123	-0.019	-0.037	0.005
low risk	0.169	0.113	-0.152	-0.041	0.006	0.085	-0.185	-0.039	0.013	0.093	-0.160	-0.040	0.094	0.094	-0.145	-0.039
momentum	0.325	0.342	-0.058	-0.011	0.374	0.358	-0.089	-0.011	0.373	0.367	-0.113	-0.014	0.465	0.366	-0.084	-0.012
profitability	0.160	0.171	-0.026	-0.008	0.165	0.178	-0.056	-0.009	0.203	0.184	-0.021	-0.008	0.336	0.203	0.002	-0.007
profit growth	0.222	0.210	0.006	0.002	0.228	0.211	0.003	0.001	0.143	0.189	0.000	0.001	0.294	0.221	-0.001	0.001
quality	0.178	0.228	-0.065	-0.007	0.206	0.242	-0.089	-0.007	0.113	0.220	-0.051	-0.005	0.265	0.248	-0.066	-0.007
seasonality	0.076	0.119	0.003	-0.003	0.095	0.123	0.006	-0.003	0.034	0.113	-0.013	-0.003	0.046	0.120	-0.022	-0.004
size	0.358	0.253	-0.018	0.009	0.195	0.208	0.015	0.012	0.021	0.190	-0.075	0.008	0.070	0.193	-0.034	0.009
skewness	0.035	0.175	0.019	0.002	-0.041	0.161	0.003	0.002	0.049	0.181	0.017	0.004	0.062	0.201	0.029	0.003
value	0.327	0.330	0.001	-0.012	0.391	0.346	0.021	-0.012	0.350	0.335	-0.006	-0.013	0.363	0.337	0.017	-0.011

Notes: The table reports theme-level CAPM alphas and betas by region. $\alpha^F_{c(i)}$ and $\beta^F_{c(i)}$ are frequentist estimates from the hierarchical linear model represented in 9. $\alpha^B_{c(i)}$ and $\beta^B_{c(i)}$ are posterior medians from the Bayesian hierarchical model with the same specification and prior structure given by fitting the same specification on world-aggregated Jensen et al. (2023) long-short factors excess returns. All values are in monthly percentage points.

Appendix D. Comparison of Frequentist and (Informative) Bayesian theme-leval CAPM alphas and betas by region

Table A9 compares frequentist and Bayesian theme-level CAPM alphas and betas across regions under informative global priors.

Appendix E. Region-Specific Backtest Performance

This section reports detailed backtest performance metrics by region and selection strategy.

Table A10: Americas — Risk and performance metrics by adjustment strategy and weighting scheme.

Adjustment	Scheme	Geom. Mean	Std. Dev.	Down Dev.	Sharpe	Sortino	Exp. Short	Ulcer	Martin	Alpha	Alpha t-stat
None	EW	0.2516	0.7989	0.4607	1.1061	0.5529	1.5320	1.9597	1.5620	0.2001	2.8292
	t_{α_i}	0.2723	0.8148	0.4608	1.1750	0.5980	1.5477	1.7935	1.8492	0.2268	3.1055
Bonferroni	EW	0.5866	1.8321	0.9252	1.1456	0.6518	2.9025	1.8878	3.8512	0.6237	3.6782
	t_{α_i}	0.5795	1.8269	0.9216	1.1345	0.6466	2.8829	1.8402	3.9018	0.6153	3.6387
BH	EW	0.4375	1.3682	0.7181	1.1349	0.6222	2.3003	1.6539	3.2521	0.4296	3.3938
	t_{α_i}	0.4502	1.3714	0.7178	1.1657	0.6401	2.3161	1.5991	3.4633	0.4471	3.5211
BY	EW	**0.7272**	1.8564	0.8943	**1.4126**	**0.8321**	2.8924	1.7771	5.1120	**0.8425**	**3.9538**
	t_{α_i}	0.7169	1.8595	0.8958	1.3894	0.8192	2.8815	1.7357	**5.1564**	0.8279	3.8745
Hochberg	EW	0.5866	1.8321	0.9252	1.1456	0.6518	2.9025	1.8878	3.8512	0.6237	3.6782
	t_{α_i}	0.5795	1.8269	0.9216	1.1345	0.6466	2.8829	1.8402	3.9018	0.6153	3.6387
Hommel	EW	0.5866	1.8321	0.9252	1.1456	0.6518	2.9025	1.8878	3.8512	0.6237	3.6782
	t_{α_i}	0.5795	1.8269	0.9216	1.1345	0.6466	2.8829	1.8402	3.9018	0.6153	3.6387
Holm	EW	0.5866	1.8321	0.9252	1.1456	0.6518	2.9025	1.8878	3.8512	0.6237	3.6782
	t_{α_i}	0.5795	1.8269	0.9216	1.1345	0.6466	2.8829	1.8402	3.9018	0.6153	3.6387
Bayesian (Conservative)	EW	0.2377	0.7721	0.4253	1.0806	0.5658	1.3737	2.0077	1.4396	0.1942	2.8077
	$t^{\mathrm{B}}_{\alpha_i}$	0.2729	0.7586	0.3915	1.2652	0.7044	1.2562	1.4132	2.3527	0.2395	3.4801
Bayesian (Global)	EW	0.1455	**0.5307**	0.3065	0.9573	0.4792	**0.9791**	1.3352	1.3180	0.1149	2.4187
	$t^{\mathrm{B}}_{\alpha_i}$	0.1940	0.5605	**0.3054**	1.2121	0.6405	0.9994	**1.0658**	2.2082	0.1671	3.2959

For each region, portfolios are constructed using equal weights (EW) or weights proportional to the frequentist t_{α_i} or posterior $t^{\mathrm{B}}_{\alpha_i}$. Selection strategies include no p-value adjustment (None), Benjamini–Hochberg (BH), Benjamini–Yekutieli (BY), Bonferroni, Hochberg, Hommel, Holm, Bayesian with a conservative prior regime, and Bayesian with an informative global prior regime. All values are reported in monthly percentage points. Best-performing strategies for each metric are shown in **bold**.

Table A11: APAC — Risk and performance metrics by adjustment strategy and weighting scheme.

Adjustment	Scheme	Geom. Mean	Std. Dev.	Down Dev.	Sharpe	Sortino	Exp. Short	Ulcer	Martin	Alpha	Alpha t-stat
None	EW	0.2016	0.5121	0.3054	1.3789	0.6643	1.4118	0.9190	2.6614	0.1544	3.9797
	t_{α_i}	0.2026	0.5352	0.3160	1.3260	0.6456	1.4072	0.9374	2.6225	0.1597	3.8159
Bonferroni	EW	0.2211	0.6553	0.3759	1.1832	0.5939	1.3441	1.8075	1.4861	0.1908	3.5049
	t_{α_i}	0.2216	0.6920	0.3971	1.1231	0.5641	1.3717	1.9962	1.3487	0.1937	3.3468
BH	EW	0.1924	0.5242	0.3187	1.2848	0.6080	1.5056	0.9604	2.4291	0.1472	3.6458
	t_{α_i}	0.1973	0.5458	0.3262	1.2662	0.6096	1.4553	0.9697	2.4687	0.1556	3.6172
BY	EW	0.2218	0.5731	0.3377	1.3573	0.6617	1.5166	0.9139	2.9483	0.1824	3.9713
	t_{α_i}	0.2236	0.6057	0.3515	1.2946	0.6412	1.4457	0.9636	2.8190	0.1873	3.7976
Hochberg	EW	**0.2270**	0.6578	0.3809	1.2102	0.6015	1.4587	1.4577	1.8919	0.1930	3.5576
	t_{α_i}	0.2259	0.6916	0.3990	1.1455	0.5721	1.4393	1.7285	1.5879	**0.1950**	3.3887
Hommel	EW	0.2231	0.6587	0.3874	1.1881	0.5816	1.5179	1.4639	1.8518	0.1890	3.4814
	t_{α_i}	0.2229	0.6907	0.4030	1.1316	0.5590	1.4789	1.7268	1.5681	0.1919	3.3400
Holm	EW	**0.2270**	0.6578	0.3809	1.2102	0.6015	1.4587	1.4577	1.8919	0.1929	3.5576
	t_{α_i}	0.2259	0.6916	0.3990	1.1455	0.5721	1.4393	1.7285	1.5879	**0.1950**	3.3887
Bayesian (Conservative)	EW	0.2090	0.5254	0.3124	1.3936	0.6732	1.3354	0.9811	2.5852	0.1562	4.0277
	$t^{\mathrm{B}}_{\alpha_i}$	0.2146	0.5646	0.3379	1.3321	0.6397	1.4915	1.0343	2.5189	0.1633	3.8055
Bayesian (Global)	EW	0.2147	**0.4855**	**0.2744**	**1.5504**	**0.7869**	**1.1058**	**0.6995**	**3.7275**	0.1667	4.6266
	$t^{\mathrm{B}}_{\alpha_i}$	0.2204	0.5276	0.3066	1.4646	0.7234	1.3253	0.8012	3.3414	0.1729	**4.2995**

For each region, portfolios are constructed using the weighting schemes discussed in Section 3.3: equal weights (EW) or weights proportional to the frequentist t_{α_i} or posterior $t^{\mathrm{B}}_{\alpha_i}$. Selection strategies include baseline no p-value adjustment (None), Benjamini–Hochberg (BH), Benjamini–Yekutieli (BY), Bonferroni, Hommel, Holm, Bayesian with a conservative prior regime, and Bayesian with an informative global prior regime. All values are reported in monthly percentage points. Best-performing strategies for each metric are shown in **bold**.

Table A12: EMEA — Risk and performance metrics by adjustment strategy and weighting scheme.

Adjustment	Scheme	Geom. Mean	Std. Dev.	Down Dev.	Sharpe	Sortino	Exp. Short	Ulcer	Martin	Alpha	Alpha t-stat
None	EW	0.3054	0.7523	0.3936	1.4303	0.7831	1.4977	1.3866	2.6882	0.2358	3.7398
	t_{α_i}	0.3245	0.7576	0.3864	1.5107	0.8471	1.4214	1.5925	2.4894	0.2554	4.0112
Bonferroni	EW	0.4168	0.9112	0.4855	1.6216	0.8671	1.8151	3.3129	1.5450	0.3586	4.4691
	t_{α_i}	**0.4238**	0.9311	0.4969	1.6141	0.8616	1.8215	3.3622	1.5484	**0.3649**	4.4471
BH	EW	0.3167	0.8234	0.4344	1.3557	0.7367	1.6300	1.8585	2.0806	0.2429	3.5004
	t_{α_i}	0.3376	0.8150	0.4176	1.4617	0.8161	1.5014	1.9110	2.1595	0.2663	3.8594
BY	EW	0.3869	0.8272	0.4144	1.6554	**0.9419**	1.4151	2.2291	2.1279	0.3191	4.5089
	t_{α_i}	0.3958	0.8455	0.4282	1.6574	0.9325	1.4694	2.4796	1.9577	0.3300	**4.5287**
Hochberg	EW	0.4004	0.8888	0.4828	1.5955	0.8374	1.8173	3.3380	1.4715	0.3442	4.3932
	t_{α_i}	0.4103	0.9117	0.4933	1.5945	0.8400	1.8184	3.3803	1.4898	0.3532	4.3913
Hommel	EW	0.3981	0.8812	0.4737	1.5999	0.8485	1.8011	3.2847	1.4868	0.3416	4.4031
	t_{α_i}	0.4077	0.9051	0.4869	1.5960	0.8458	1.8085	3.3358	1.5001	0.3505	4.3936
Holm	EW	0.4004	0.8888	0.4828	1.5955	0.8374	1.8173	3.3380	1.4715	0.3442	4.3932
	t_{α_i}	0.4103	0.9117	0.4933	1.5945	0.8400	1.8184	3.3803	1.4898	0.3532	4.3913
Bayesian (Conservative)	EW	0.3119	0.7658	0.3938	1.4356	0.7996	1.5313	1.4821	2.5696	0.2381	3.7426
	$t^{\mathrm{B}}_{\alpha_i}$	0.3277	0.7784	0.3982	1.4852	0.8305	1.5508	1.5329	2.6123	0.2553	3.9189
Bayesian (Global)	EW	0.2818	**0.6792**	**0.3394**	1.4597	0.8370	**1.2569**	**1.1436**	3.0029	0.2164	3.8301
	$t^{\mathrm{B}}_{\alpha_i}$	0.3122	0.7015	0.3518	**1.5682**	0.8943	1.3692	1.1353	**3.3571**	0.2465	4.2000

For each region, portfolios are constructed using the weighting schemes discussed in Section 3.3: equal weights (EW) or weights proportional to the frequentist t_{α_i} or posterior $t^{\mathrm{B}}_{\alpha_i}$. Selection strategies include baseline no p-value adjustment (None), Benjamini–Hochberg (BH), Benjamini–Yekutieli (BY), Bonferroni, Hommel, Holm, Bayesian with a conservative prior regime, and Bayesian with an informative global prior regime. All values are reported in monthly percentage points. Best-performing strategies for each metric are shown in **bold**.