# Asset Heterogeneity and Uncommon Factors *

Lin William Cong Guanhao Feng Jingyu He Junye Li Qianshu Zhang

First version: Sep. 2022; this version: January 15, 2026

**Abstract**

Standard factor models typically rely on a "one-size-fits-all" assumption, overlooking the fact that pricing relationships vary across assets. This paper investigates whether assets exhibit grouped heterogeneity and adhere to uncommon factor models. Unlike common factors that price the entire cross-section, uncommon factors provide significant explanatory power only within specific asset clusters or macroeconomic regimes. Empirically, we find pervasive evidence of heterogeneous risk premia among U.S. equities. We demonstrate that the "factor zoo" is neither globally sparse nor stable; instead, certain factors command risk premia only in localized subsets while remaining irrelevant elsewhere. Leveraging Bayesian asset pricing for factor selection, our model significantly improves out-of-sample predictive performance and investment returns. Our findings resolve conflicting evidence regarding factor stability and emphasize that asset pricing is fundamentally state-dependent and asset-specific.

**Key Words:** Heterogeneous Risk Premia, Grouped Heterogeneity, Factor Selection, Bayesian Asset Pricing, Macroeconomic Regimes

**JEL Classification:** C11, C38, G11, G12.

*We thank Rohit Allena (discussant), Doron Avramov, Yi Cao (discussant), I-Hsuan Ethan Chiang (discussant), Siddhartha Chib, Tarun Chordia, John Cochrane, Darrell Duffie, Jianqing Fan, Stefano Giglio, P. Richard Hahn, Cam Harvey, Zhiguo He, Yael Hochberg, Yongmiao Hong, David Hirshleifer, Bob Jarrow, Fuwei Jiang (discussant), Bryan Kelly, Serhiy Kozak, Sophia Zhengzi Li, David Ng, Andrew Patton, Markus Pelger, Xiao Qiao, Alberto Rossi, Olivier Scaillet, Michael Sockin, Oleg Sokolinskiy (discussant), Artem Streltsov, Daniel Titman, Fabio Trojani, Junbo Wang (discussant), Dacheng Xiu, Mao Ye, Guofu Zhou, and seminar and conference participants at BlackRock, CityU HK, Columbia, Cornell SC Johnson, Cornell Tech, CUHK, 2024 China Fintech Research Conference, EasternFA, 2024 EFMA Annual Meeting, 3rd Frontiers of Factor Investing Conference, Harvest Fund 25th Anniversary Ceremony, HKU, HUST, 4th International FinTech Research Forum (RUC), KAIST Digital Finance Conference, Macquarie University, MFA, Mid-South DATA Conference, NBER-NSF Seminar on Bayesian Inference in Econometrics and Statistics, ESIF Economics and AI+ML Meeting (Cornell), NYU Courant, Oxford, Peking University Guanghua, PHBS, Princeton, 2023 Tongji Finance Symposium, Rice, Rutgers Business School, 6th Shanghai Financial Forefront Symposium, Singapore National University, Stanford MS&E, SWFA, Tsinghua University SEM, UCAS, University of Geneva, USC Marshall, UT Austin, 4th Workshop on Big Data Econometric Theory and Application, Xiamen University, and 2023 XJTLU AI and Big Data in Accounting and Finance Conference for invaluable comments and discussions. We thank Ripple's UBRI for research support. Cong (Corresponding, e-mail: will.cong@cornell.edu) is at Cornell University and NBER; Feng (E-mail: gavin.feng@cityu.edu.hk), He (E-mail: jingyuhe@cityu.edu.hk) and Zhang (E-mail: qszhang7-c@my.cityu.edu.hk) are at City University of Hong Kong; Li (E-mail: li_junye@fudan.edu.cn) is at Fudan University.

# 1 Introduction

Traditional implementations of common factor models typically rely on a restrictive "one-size-fits-all" assumption, estimating a single set of factor risk premia across the entire cross-section of assets. While recent literature has responded to the burgeoning "factor zoo" by seeking a globally sparse set of universal factors (e.g., Feng, Giglio, and Xiu, 2020; Bryzgalova, Huang, and Julliard, 2023), this approach implicitly assumes that pricing relationships are invariant across the market. This assumption sits at the heart of an unresolved debate: whether the stochastic discount factor is truly sparse or inherently dense—a phenomenon some describe as an "illusion of sparsity" (Kozak et al., 2020; Giannone et al., 2021).

However, such models often overlook heterogeneous risk premia, wherein specific factors command significant premia only within distinct asset groups while remaining irrelevant elsewhere, effectively acting as "uncommon factors." This makes it particularly challenging to identify and estimate a parsimonious *universal* model with common factors that accurately explains all return variation, especially when focusing on individual assets rather than portfolios. By forcing a universal model onto a segmented market, researchers risk discarding locally significant signals as noise and overlooking pervasive risk price variation (Patton and Weller, 2022).

In this paper, we challenge the universal pricing paradigm by investigating whether observed asset heterogeneity is driven by these heterogeneous risk premia and their adherence to uncommon factor models. This conceptual shift recognizes that the "factor zoo" is neither globally sparse nor stable; instead, priced risk factors rotate dynamically in response to firm characteristics or the broader economic state (Smith and Timmermann, 2022; Cui et al., 2025). The goal of uncommon factor models is to complement, rather than replace, common factor models.

Preliminary empirical evidence based on clustering stocks simply by market equity or the federal funds rate indicates that heterogeneity in risk premia do exist in the cross-section and time-series. However, given the large number of characteristics

and macroeconomic variables, it is not clear which underlying characteristics are responsible for this heterogeneity and which factors are relevant for which groups of assets. These considerations point to the need for a unified framework that can jointly uncover asset groupings and select the appropriate factors within each group in a systematic way.

We address the issue by integrating the Bayesian Clustering Model (BCM, Cong, Feng, He, and Li, 2024) with period-by-period Fama-MacBeth cross-sectional regressions, which we term the Bayesian Fama-Macbeth Clustering Model (BFCM). Clustering aims to group asset-return observations for cluster-wise cross-sectional factor models. Specifically, BFCM tackles the econometric challenge by simultaneously selecting factors on the right-hand side and clustering observations on the left-hand side of regressions. It generates asset return clusters with economic definitions based on firm characteristics or macroeconomic variables. Our approach provides a novel, interpretable, and computationally efficient method for analyzing unbalanced panel data with grouped heterogeneity, enabling the discovery of uncommon factors with heterogeneous risk premia.

We first examine asset heterogeneity and uncommon factors in the cross section of stock returns. We consider the entire universe of U.S. individual stocks during the sample period from 1980 to 2024 with 20 highly cited firm characteristics from the literature. Return variance, market equity, and earnings-to-price ratios are the most important characteristics describing the grouped heterogeneity in characteristics-managed clusters. Aside for the Market factor (the intercept in cross-sectional regression), the ME (Size factor) is the most common factor, demonstrating high selection probabilities across most clusters.[1]

Regarding predictive power, the BFCM clustered model exhibits improved performance, particularly in terms of the predictive $R^2$ within low-volatility (SVAR) clusters. This improvement can be translated into an investment strategy by forming long-short (20%) portfolios within each cluster. The cluster-based long-short portfolio

[1]The list of characteristics and their acronyms is presented in Section 2.1.

significantly outperforms the overall portfolio without clustering, particularly in the value-weighted case.

We further expand the heterogeneity to time series by splitting based on macroeconomic variables. We identify significant time-varying heterogeneity driven by aggregate liquidity and net equity issuance, revealing that risk premia are highly state-dependent rather than static. These findings confirm that the "Factor Zoo" is dynamic in both cross-section and time-series, with the factor risk premia shifting endogenously across both asset clusters and macroeconomic states.

**Literature.** Our paper makes a direct contribution to a substantial literature on factor model selection. However, most of them treat test assets arbitrarily. We suggest evaluating linear factor models by jointly selecting factors and test assets and modeling grouped heterogeneity (uncommon factor selection) to complement existing studies in empirical asset pricing. The literature has recognized heterogeneous factor structures in asset returns; however, almost all analyses restrict themselves to assets exogenously grouped according to heuristics, statistical patterns, or researchers' *a priori* knowledge. For example, factor models are developed specifically for assets in a particular country, an asset class, or a group with strong co-movements (see, e.g., Foerster and Karolyi, 1999; Griffin, 2002; Karolyi and Stulz, 2003; Bekaert, Hodrick, and Zhang, 2009; Hou, Karolyi, and Kho, 2011; Jarrow, Murataj, Wells, and Zhu, 2020; Chaieb, Langlois, and Scaillet, 2021).

Closely related to ours by also analyzing endogenous grouped heterogeneity in financial markets, Ahn, Conrad, and Dittmar (2009) use unsupervised clustering based on return correlations, and Patton and Weller (2022) build on Bonhomme and Manresa (2015) to generalize $K$-means to group assets based on within-group slopes and averages, and find risk-price heterogeneity pervasive and important. Our paper distinguishes itself by emphasizing interpretability and jointly selecting variables while clustering observations in the panel data.

In terms of methodology, BFCM integrates decision trees and Bayesian statistical methods into a unified framework for sparse modeling with grouped heterogeneity.

Bayesian methods have long been applied in empirical finance (see, e.g., Avramov and Chordia, 2006; Barillas and Shanken, 2018; Chib and Zeng, 2020; Avramov et al., 2023; Chib et al., 2023), and are argued to be the ultimate solution to asset pricing (see the AFA presidential address by Harvey, 2017). We use spike-and-slab priors to estimate the marginal likelihood of asset returns, rather than identifying common linear factors or estimating risk premia for fixed test assets. BFCM computes efficiently and offers insights and performance improvement in pricing and investments by combining information on characteristics and factors, rather than enumerating all variable combinations. Two recent studies by Hwang and Rubesam (2020) and Bryzgalova, Huang, and Julliard (2023) apply Bayesian variable selection methods to factor model selection.

Tree-based machine learning is gaining attention in asset pricing due to its ability to graphically represent nonlinear interactions and effectively model high-dimensional characteristics (e.g., Rossi and Timmermann, 2015; Gu et al., 2020; Bianchi et al., 2021; Van Binsbergen et al., 2023). Recently, Bryzgalova et al. (2025) apply shrinkage to prune a forest of shallow trees to select effective basis portfolios. Our paper, along with Cong et al. (2025), is among the first to develop goal-oriented clustering, which searches in a flexible clustering space in a data-driven and yet economically guided manner. Cong et al. (2025) introduce "panel trees" to generalize security sorting for deriving basis portfolios and common latent factors. Relative to ensemble models (XBART, He et al., 2019; He and Hahn, 2021), BFCM focuses on clustering using natural global split criteria that prevent overfitting and preserve single-tree interpretability.

The remainder of the article is organized as follows. Section 2 describes the data and preliminary evidence. Section 3 presents the model to find economic-driven heterogeneity. Section 4 presents the baseline empirical findings of group heterogeneity in the cross section of stock returns. Section 5 extends the heterogeneity to both time series and cross-sections. Section 6 concludes.

## 2 Data and Preliminary Evidence

### 2.1 Data

**Assets.** We consider the entire universe of U.S. individual equities. Our sample spans from January 1980 to December 2024.[2] The first 30 years of the sample are used for model estimation, while the most recent 14 years are for out-of-sample tests in the main empirical section. In the training sample, the average and median number of stocks are 4,996 and 4,780, respectively. In the testing sample, these numbers are 3,608 and 3,588.

**Characteristics.** We use 20 firm characteristics with monthly observations for each asset: market equity (ME), market beta (BETA), bid-ask spread (BAS), stock CAPM residual variance (SVAR), book-to-market (BM), earnings-to-price (EP), cash flows-to-price (CFP), sales-to-price (SP), asset growth (AGR), net equity issuance (NI), accounting accruals (ACC), operating profitability (OP), return on equity (ROE), momentum (MOM), short-term reversal (STR), seasonality (SEAS), advertisement-to-market (ADM), R&D-to-market (RDM), unexpected quarterly earnings (SUE), and earnings-announcement abnormal return (ABR).[3] We rank each characteristic in each month cross-sectionally and standardize the ranks to a [-1, 1] scale. Missing values are imputed to 0. Quintiles are used as split value candidates for each characteristic, mimicking the conventional $5 \times 1$ sorting in asset pricing. For 20 characteristics, this results in 80 ($4 \times 20$) splitting candidates. Note that these characteristics are not only used for cluster splitting but also for CS factor loadings. When used for CS factor regression, the characteristics are standardized to have mean 0 and standard deviation 1 following Fama and French (2020).

**Macroeconomic variables.** Following Welch and Goyal (2008), we incorporate ten macroeconomic variables to detect time-series breaks and regimes relevant to asset

[2]We apply standard filters (see, e.g., Fama and French, 2015): (1) including only stocks listed on NYSE, AMEX, or NASDAQ for more than one year, (2) using firms with CRSP share codes of 10 and 11, and (3) excluding stocks with negative book equity or negative lagged market equity.

[3]These characteristics cover both firm fundamentals and market signals, span all commonly used categories in the literature, and were mostly introduced before the test sample period.

pricing. These variables include the 3-month Treasury bill rate, inflation, term spread, default spread, dividend yield, earnings-to-price, market volatility, net equity issues, leverage, and liquidity. We adjust the macro indicators by subtracting the previous 10-year rolling window average and then apply a 12-month smoothing to reduce volatility. If a macro indicator is greater than 0, it means the current level is higher than its 10-year average, indicating a high level of economic activity. This approach ensures the consistent evaluation of each predictor's performance in the tree clustering method for detecting macroeconomic regimes using recent data.

### 2.2 Preliminary Evidence of Heterogeneity in the Cross-section

Before deploying the comprehensive framework to endogenously discover asset clusters, we first motivate our approach by examining whether risk premia are indeed heterogeneous across simple groupings, for example, by market equity. If the "one-size-fits-all" assumption of standard cross-sectional factor models holds, the estimated risk premia for a given factor should be statistically indistinguishable across different subsets of the asset universe.

To test this, we partition the universe of U.S. individual stocks into three intuitive groups based on market equity (ME): "Small" (bottom 33%), "Mid" (between 33% to 66%) and "Big" stocks (top 33%). We then estimate the risk premia of cross-sectional (CS) factors Fama and French (2020) within each group separately using standard Fama-Macbeth regressions and compare them to the estimates obtained from the overall pooled sample. Table 1 presents the preliminary evidence. Panel A reports the risk premia estimates for the Small, Mid, Big, and All groups. Panel B provides the $p$-values of the two-sample test for tests of equality of risk premia between these groups.

The results reveal striking heterogeneity. For liquidity-related factors such as Short-Term Reversal (STR) and Bid-Ask Spread (BAS), the $p$-values for the differences between Small and Big stocks are effectively zero ($0.00$). Similarly, for fundamental characteritics like Book-to-Market (BM) and Standardized Unexpected Earnings (SUE), we strongly reject equality between the Small and Mid-cap stocks ($p = 0.00$ for

Table 1: **Pre-Evidence: Test of Equality of Risk Premia between Size Groups**

Panel A of this table presents the OLS Fama-Macbeth estimate of factor risk premia for each cluster, with the columns representing small, mid, big market equity clusters and all pooled data. Panel B gives the $p$-value of the frequentist test of equality of risk premia between each group and the overall. The sample period is based on the full sample (1980-2024).

| | Small | Mid | Big | All | Small-Mid | Small-Big | Mid-Big | Small-All | Mid-All | Big-All |
|---|---|---|---|---|---|---|---|---|---|---|
| | Panel A: Risk Premia | | | | Panel B: $p$-value | | | | | |
| ME | -1.07 | -0.01 | -0.44 | -0.18 | 0.00 | 0.00 | 0.00 | 0.00 | 0.14 | 0.03 |
| ABR | 0.25 | 0.21 | 0.10 | 0.22 | 0.17 | 0.00 | 0.00 | 0.27 | 0.62 | 0.00 |
| SUE | 1.41 | 0.78 | 0.32 | 0.83 | 0.00 | 0.00 | 0.00 | 0.00 | 0.30 | 0.00 |
| BAS | -0.22 | -0.07 | -0.07 | -0.12 | 0.00 | 0.00 | 0.83 | 0.01 | 0.21 | 0.15 |
| MOM | 0.01 | 0.18 | 0.16 | 0.12 | 0.03 | 0.07 | 0.86 | 0.15 | 0.48 | 0.64 |
| BM | 0.43 | 0.07 | 0.06 | 0.26 | 0.00 | 0.00 | 0.85 | 0.01 | 0.00 | 0.00 |
| EP | 0.13 | 0.02 | 0.03 | 0.01 | 0.16 | 0.23 | 0.93 | 0.10 | 0.89 | 0.82 |
| CFP | 0.16 | 0.13 | -0.02 | 0.08 | 0.78 | 0.04 | 0.07 | 0.31 | 0.47 | 0.19 |
| SP | 0.21 | 0.19 | 0.14 | 0.18 | 0.80 | 0.29 | 0.44 | 0.66 | 0.87 | 0.50 |
| AGR | -0.17 | -0.13 | -0.08 | -0.13 | 0.32 | 0.03 | 0.17 | 0.24 | 0.87 | 0.19 |
| NI | -0.22 | -0.02 | -0.07 | -0.11 | 0.00 | 0.00 | 0.16 | 0.01 | 0.00 | 0.09 |
| ACC | -0.13 | -0.16 | -0.19 | -0.16 | 0.54 | 0.15 | 0.39 | 0.40 | 0.85 | 0.45 |
| OP | 0.17 | 0.04 | 0.10 | 0.17 | 0.02 | 0.24 | 0.29 | 0.91 | 0.01 | 0.21 |
| ROE | 0.10 | 0.11 | 0.07 | 0.09 | 0.80 | 0.57 | 0.38 | 0.90 | 0.66 | 0.57 |
| SEAS | 0.14 | 0.02 | 0.05 | 0.08 | 0.00 | 0.05 | 0.46 | 0.16 | 0.09 | 0.43 |
| ADM | -0.02 | 0.02 | -0.01 | -0.01 | 0.32 | 0.82 | 0.40 | 0.79 | 0.36 | 0.99 |
| RDM | 0.37 | 0.26 | 0.09 | 0.25 | 0.03 | 0.00 | 0.00 | 0.01 | 0.74 | 0.00 |
| SVAR | -0.18 | -0.34 | -0.15 | -0.21 | 0.09 | 0.81 | 0.07 | 0.75 | 0.15 | 0.59 |
| BETA | -0.01 | 0.07 | 0.08 | 0.08 | 0.27 | 0.26 | 0.89 | 0.20 | 0.93 | 0.96 |
| STR | -0.86 | -0.43 | -0.34 | -0.58 | 0.00 | 0.00 | 0.19 | 0.00 | 0.02 | 0.00 |

both). The Net Equity Issuance (NI) premium is substantially more negative for small stocks than for mid or large stocks, and equality is rejected in most relevant pairwise comparisons, further confirming the evidence.

These findings strongly reject the null hypothesis of universal risk premia. A pooled model that forces a single risk premium across all assets would mask these nuances, likely underestimating risks for small-cap stocks and overestimating them for large stocks. The significant heterogeneity observed in this simple size-based cluster motivates the need for a more sophisticated approach to identify the optimal partitions of the asset universe, which we address in the subsequent sections.

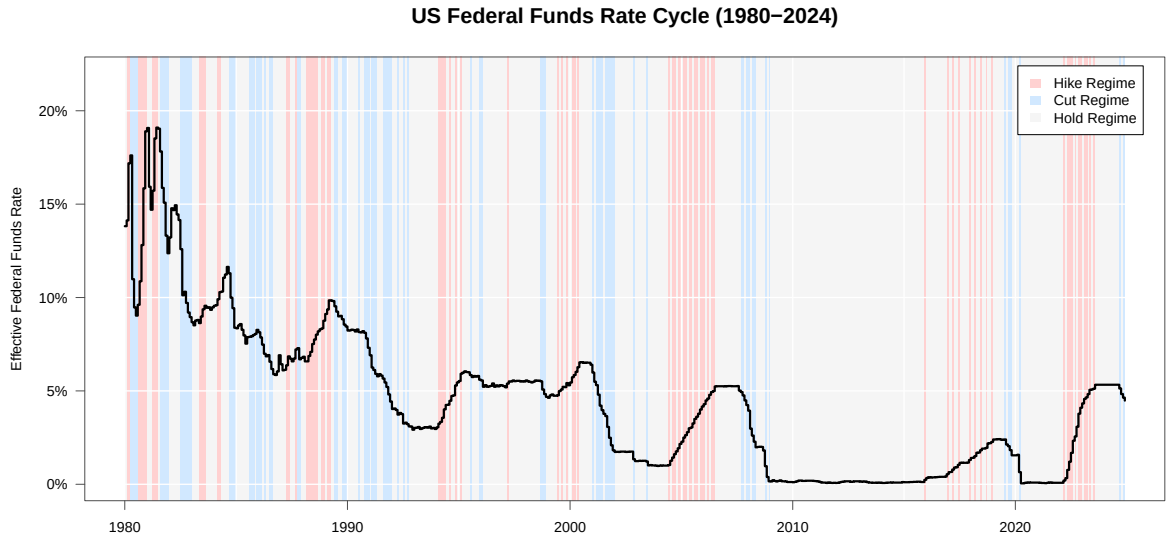
### 2.3 Preliminary Evidence of Heterogeneity in the Time Series

We next examine the evidence of stability of risk premia over time. Standard unconditional factor models typically estimate a single average premium for the full sample, implicitly assuming that the compensation for risk is constant regardless of the

macroeconomic state. To investigate this, we partition our sample into three regimes based on the Federal fund effective rate decisions—Hike (rising rates), Cut (falling rates), and Hold (stable rates)—which serve as proxies for the broader economic cycle and liquidity environment. Figure 1 plot the time-series of fund rate and regimes featured by different colors. We then estimate the risk premia for the 20 characteristics within each regime using OLS Fama-MacBeth regressions.

Figure 1: Fed Fund Rate Regimes

This figure plots the US Federal fund rate (from FRED), with different color indicate different regime (Hike, Cut or Hold of Fund rate).




The results, presented in Table 2, demonstrate that risk premia are unstable, with many factors exhibiting pronounced pro-cyclical or counter-cyclical behavior. When comparing Hike versus Cut regimes, we observe $p$-values of $0.00$ for Earnings-to-Price (EP), Short-Term Reversal (STR), and Cash Flow-to-Price (CFP). The rejection of equality for STR is particularly notable; the premium shifts from $-0.39$ in Hikes to $-0.97$ in Cuts, with a $p$-value of $0.00$. Similarly, the sign flip in Earnings-to-Price—from significantly negative in Hikes to significantly positive in Cuts—is statistically validated by a $p$-value of $0.00$.

Although the statistical evidence for heterogeneity across these broad monetary regimes is arguably less ubiquitous than the stark disparities observed in the cross-section, the economic implications of these shifts remain profound. The instability

observed across these major factors underscores the limitations of the assumption of homogeneous risk premia in the cross-section and time-series and motivates the application of a more sophisticated model to endogenously identify economic-driven clusters, as detailed in the following sections.

Table 2: Pre-Evidence of Time-varying Risk Premia

Panel A of this table presents the OLS Fama-Macbeth estimate of risk premia, with the column matching the time period for federal fund rate hike, cut and hold. Panel B gives the $p$-value of frequentist test of equality of risk premia between this two clusters. The sample period is based on the full sample (1980-2024).

| | Hike | Cut | Hold | All | Hike-Cut | Hike-Hold | Cut-Hold | Hike-All | Cut-All | Hold-All |
|---|---|---|---|---|---|---|---|---|---|---|
| | Panel A: Risk Premia | | | | Panel B: $p$-value | | | | | |
| ME | -0.22 | -0.02 | -0.20 | -0.18 | 0.36 | 0.93 | 0.31 | 0.79 | 0.38 | 0.77 |
| ABR | 0.18 | 0.27 | 0.22 | 0.22 | 0.10 | 0.31 | 0.27 | 0.27 | 0.28 | 0.94 |
| SUE | 0.76 | 0.85 | 0.83 | 0.83 | 0.30 | 0.26 | 0.86 | 0.30 | 0.75 | 0.84 |
| BAS | 0.04 | -0.10 | -0.16 | -0.12 | 0.10 | 0.01 | 0.36 | 0.02 | 0.77 | 0.28 |
| MOM | 0.29 | 0.06 | 0.10 | 0.12 | 0.25 | 0.11 | 0.84 | 0.14 | 0.74 | 0.80 |
| BM | 0.26 | 0.34 | 0.25 | 0.26 | 0.63 | 0.95 | 0.49 | 0.96 | 0.56 | 0.81 |
| EP | -0.17 | 0.50 | -0.05 | 0.01 | 0.00 | 0.31 | 0.00 | 0.11 | 0.00 | 0.35 |
| CFP | 0.32 | -0.25 | 0.10 | 0.08 | 0.00 | 0.07 | 0.02 | 0.04 | 0.02 | 0.77 |
| SP | -0.03 | 0.00 | 0.27 | 0.18 | 0.86 | 0.01 | 0.03 | 0.05 | 0.14 | 0.19 |
| AGR | -0.07 | -0.15 | -0.13 | -0.13 | 0.28 | 0.25 | 0.78 | 0.28 | 0.70 | 0.84 |
| NI | -0.11 | -0.06 | -0.12 | -0.11 | 0.41 | 0.85 | 0.19 | 0.99 | 0.26 | 0.73 |
| ACC | -0.14 | -0.11 | -0.18 | -0.16 | 0.73 | 0.47 | 0.30 | 0.66 | 0.43 | 0.61 |
| OP | 0.20 | 0.16 | 0.16 | 0.17 | 0.76 | 0.71 | 0.99 | 0.75 | 0.94 | 0.91 |
| ROE | 0.08 | 0.14 | 0.09 | 0.09 | 0.48 | 0.98 | 0.41 | 0.86 | 0.46 | 0.82 |
| SEAS | 0.06 | 0.11 | 0.08 | 0.08 | 0.57 | 0.77 | 0.66 | 0.74 | 0.67 | 0.95 |
| ADM | -0.04 | -0.02 | 0.00 | -0.01 | 0.82 | 0.41 | 0.61 | 0.52 | 0.75 | 0.71 |
| RDM | 0.26 | 0.30 | 0.24 | 0.25 | 0.71 | 0.70 | 0.37 | 0.84 | 0.47 | 0.74 |
| SVAR | -0.13 | -0.34 | -0.20 | -0.21 | 0.40 | 0.73 | 0.46 | 0.68 | 0.49 | 0.91 |
| BETA | -0.06 | 0.36 | 0.04 | 0.08 | 0.02 | 0.41 | 0.03 | 0.26 | 0.05 | 0.68 |
| STR | -0.39 | -0.97 | -0.54 | -0.58 | 0.00 | 0.18 | 0.01 | 0.08 | 0.03 | 0.53 |

## 3 Methodology

While the preliminary evidence confirms that risk premia are neither uniform across assets nor stable over time, these simple splits based on market equity and federal fund rate regimes represent only a fraction of the potential sources of heterogeneity. In reality, the conditioning information set is vast, comprising dozens of firm characteristics and macroeconomic indicators that could plausibly define distinct pricing clusters or economic states. Manually testing every possible permutation of characteristics and macro regimes is infeasible. Furthermore, factors' premia vary across different subsets, arising the problem of factor selecion — which factor is specific to

which cluster is unknown. Thus, we require a framework that can perform heterogeneous clustering with cluster specific factor selection simultaneously.

Cong et al. (2024) propose the Bayesian Clustering Model (BCM) to perform clustering and factor loading selection in financial panel data. We extend it to the context of heterogeneous factor risk premia through period-by-period Fama-Macbeth cross-sectional regression. Section 3.1 introduces the cross-sectional regression model of asset returns, Section 3.2 discusses the factor risk premia selection in each cluster using Bayesian Spike-and-Slab priors, and then Section 3.3 presents the algorithm for clustering observations and fitting a heterogeneous, uncommon factor model for each leaf cluster.

### 3.1 Uncommon Cross-Sectional Factor Model

We first introduce the complete model of the data-generating process, assuming asset clusters are known, and provide further details on how to find clusters in Section 3.3. Let $r_{i,t}$ denote the excess return of the individual asset $i$ at time $t$, where the index $i$ ranges from $1$ to $n_t$, the number of assets available at time $t$. We let $\mathbf{z}_{i,t-1}$ denote the vector of $K$ characteristics of the $i$-th asset at the time period $t-1$. We work with the cross-sectional (CS) factor model in Fama and French (2020), where the factor loadings are the characteristics $\mathbf{z}_{i,t-1}$ given at the time $t-1$. The period-by-period cross-sectional regression of individual stock return on lagged characteristics produced the realization of the CS factors. Our primary focus is on the risk premia of CS factors, estimated as the time-series averages of the factor returns. Suppose asset returns exhibit grouped heterogeneity in the cross section, in that they follow a potentially different factor structure (as reflected in the risk premia on the CS factors)

across the groups:

$$
\begin{aligned}
r_{i,t} &= A(\mathbf{z}_{i,t-1}) + \mathbf{z}_{i,t-1}^{\mathsf{T}} F(\mathbf{z}_{i,t-1}) + \epsilon_{i,t}, \\
\textit{where} \qquad A(\mathbf{z}_{i,t-1}) &= \sum_{j=1}^{J} \mathbb{1}_{\{T(\mathbf{z}_{i,t-1})=j\}} r_{j,z,t}, \\
F(\mathbf{z}_{i,t-1}) &= \sum_{j=1}^{J} \mathbb{1}_{\{T(\mathbf{z}_{i,t-1})=j\}} \mathbf{f}_{j,t}, \\
\epsilon_{i,t} \overset{\text{independent}}{\sim} N(0, \sigma_t^2(\mathbf{z}_{i,t-1})), \quad \sigma_t^2(\mathbf{z}_{i,t-1}) &= \sum_{j=1}^{J} \mathbb{1}_{\{T(\mathbf{z}_{i,t-1})=j\}} \sigma_{j,t}^2,
\end{aligned}
\tag{1}
$$

where $\mathbf{f}_{j,t}$ is a vector of $K$ estimated CS factors, $I_K$, a $K \times K$ identity matrix, $\mathbb{1}(\cdot)$, the indicator function, and $\otimes$, the Kronecker product. The clustering function $T(\mathbf{z}_{i,t-1})$ describes the latent grouped heterogeneity, which is assumed to be a deterministic function of asset characteristics $\mathbf{z}_{i,t-1}$. It maps the characteristics for each unit observation to one (and only one) group label in $1, 2, \ldots, J$, with $J$ clusters in total. For simplicity, we specify $A(\cdot)$ as a cluster-specific scalar, and $F(\cdot)$ as a $K \times 1$ vector for heterogeneous level return and CS factors, respectively.

The clustering function $T(\cdot)$ is empirically learned through a customized tree that partitions the asset space based on firm characteristics (or macroeconomic variables) through a sequence of split rules (cutpoints). In the main empirical part, the BFCM tree is driven solely by firm characteristics $T(\mathbf{z}_{i,t-1})$ to identify the cross-sectional heterogeneity of asset returns; then we further allow the tree to be driven by both firm characteristics and macro predictors $T(\mathbf{z}_{i,t-1}, \mathbf{x}_{t-1})$. This extension enables the model to capture heterogeneity along both the cross-sectional and time-series dimensions with regime shifts.

### 3.2 Factor Risk Premia Selection

We begin with Bayesian model selection within a given cluster (a single leaf node of the tree), focusing on asset return observations, $r_{i,t}$, within the same $j$-th cluster for some $j$. Without worrying about cluster assignment, substituting the dynamics of $A(\cdot)$

and $F(\cdot)$ into the return dynamics, one gets:

$$r_{i,t} = r_{j,z,t} + \mathbf{z}_{i,t-1}^{\mathsf{T}} \mathbf{f}_{j,t} + \epsilon_{i,t}. \tag{2}$$

The coefficients vector $\mathbf{f}_{j,t} = \{f_{j,t,k} \mid 1 \leq k \leq K\}$ has length $K$ and represent the CS factors in cluster $j$ at time $t$.

**Spike-and-Slab prior.** Notice that Equation (2) consists of factor $\mathbf{f}_t$ that are results of the dynamic slope at time $t$. The number of factors and firm characteristics can be large, resulting in a large number of variables in the model. We perform Bayesian variable selection on CS factor risk premia. In other words, we do not determine the usefulness of a factor for each specific time period; instead, we evaluate its overall contribution across all time periods. The selection of a factor is based on the accumulation of evidence.

We assume independent priors for the regression coefficients vector $\mathbf{f}_j$. A vector of the latent variables taking the value of 1 or 0 to indicate whether the prior on the corresponding coefficient $\mathbf{f}_j$ is in the region of "slab" or "spike," i.e., whether the corresponding variable in (2) is selected in the model or not,

$$\boldsymbol{\gamma}_j = (\gamma_{j,1}, \gamma_{j,2}, \cdots, \gamma_{j,K}) = (\underbrace{\boldsymbol{\gamma}_{j,\mathbf{f}}}_{K \times 1}). \tag{3}$$

The prior distributions of the regression coefficients are:

$$\begin{aligned}
&\pi(f_{j,k,t} \mid \sigma_{j,t}^2, \boldsymbol{\gamma}_j) = (1 - \gamma_{j,\mathbf{f},k}) N(0, \xi_0^2 \sigma_j^2) + \gamma_{j,\mathbf{f},k} N(0, \xi_1^2 \sigma_j^2), \\
&\pi(r_{j,z,t} \mid \sigma_{j,t}^2) = N(0, \xi^2 \sigma_{j,t}^2), \\
&\pi(\sigma_j^2) = \text{inverse-Gamma}(S_0, \upsilon_0), \\
&\pi(\boldsymbol{\gamma}_j) = \pi(\boldsymbol{\gamma}_{j,\mathbf{f}}) = \prod_{k=1}^{K} w_k^{\gamma_{j,\mathbf{f},k}} (1 - w_k)^{(1-\gamma_{j,\mathbf{f},k})},
\end{aligned} \tag{4}$$

for $k = 1, \cdots, K$, and $i = 1, \cdots, M$. We employ the continuous version of the spike-and-slab prior for factor selection (George and McCulloch, 1993), which is a mixture

of two normal distributions with different variances. If $\gamma_{j,\mathbf{f},k} = 1$, the corresponding $k$-th factor $f_{t,k}$ is selected, and the prior is in the "slab" with large $\xi_1$. Thus, the coefficient is less shrunk. If $\gamma_{j,\mathbf{f},k} = 0$, the $k$-th factor is not selected, and the prior is in the "spike" with very small $\xi_0$ to shrink the coefficients towards zero. In addition, $\xi$ is large to reflect almost no shrinkage of the level return $r_{j,z,t}$. Note that factor selection is performed simultaneously across the time series: evidence for including a factor is assessed over each sample period t.

We choose prior hyperparameter $w_k = 0.5$, implying an equal prior probability to select or remove a factor. Furthermore, we assume the residual variance follows a standard conjugate Inverse-Gamma prior distribution.

### 3.3 Clustering through BFCM

With the knowledge of how we select factors (and models) in each cluster, we discuss how the clusters are jointly determined. Our clustering algorithm differs from Cong et al. (2024) in the splitting criterion. A detailed methodology is provided in the Internet Appendix II.1. We only discuss the splitting criterion below.

The partitioning generally aims to group similar observations into the same leaf to fit a locally sparse model well. To evaluate split rule candidates, the "fitness" of the resulting factor model at each leaf is a natural split criterion. However, the commonly used goodness of fitness, likelihood function, of the model in (2) involves unknown parameters that cannot be accurately estimated given the noisy data, which may favor a bad split rule. Instead, we use the closed-form expression of the *marginal likelihood*, where all fitted parameters are integrated out, to address any concerns about parameter uncertainty during tree growth.

**Marginal likelihood.** At each month $t$, stack all data in the *same cluster* in matrix form, $\mathbf{R}_t = [r_{1,t}, ..., r_{n,t}]^\mathsf{T}$, $\mathbf{Z}_{t-1} = [\mathbf{z}_{1,t-1}, ..., \mathbf{z}_{n,t-1}]^\mathsf{T}$, then the marginal likelihood of the model

at node $\mathcal{A}_0$ for month $t$ is given by:

$$p(\mathcal{A}_0) := \prod_{t=1}^{T} p(\mathbf{R}_t \mid \mathbf{Z}_{t-1}) = \prod_{t=1}^{T} \int p(\mathbf{R}_t \mid \mathbf{Z}_{t-1}, \boldsymbol{\gamma}_j, r_{j,z,t}, \mathbf{f}_j, \sigma_j^2) \\ \times \pi(r_{j,z,t} \mid \sigma_{j,t}^2)\pi(\mathbf{f}_{j,t} \mid \sigma_j^2, \boldsymbol{\gamma}_j)\pi(\sigma_{j,t}^2 \mid \boldsymbol{\gamma}_j)\pi(\boldsymbol{\gamma}_j) dr_{j,z,t} d\mathbf{f}_{j,t} d\sigma_{j,t}^2 d\boldsymbol{\gamma}_j. \tag{5}$$

Intuitively, the marginal likelihood takes the expectation of the unknown parameters in the likelihood function with respect to the prior distributions. A function of the data and prior parameters only, it accounts for parameter estimation and model selection uncertainties, separating the tree growth from factor model estimations. We describe in Proposition 1 the marginal likelihood.

**Proposition 1.** *Stacking all asset returns in the node in a vector* $\mathbf{R}$*, the marginal likelihood in (5) has a closed form:*

$$p(\mathbf{R} \mid \mathbf{Z}) = \frac{1}{2^K} \sum_{I_m} \left[ \prod_{t=1}^{T} \pi(\boldsymbol{\gamma}_{j,\mathbf{f}} = I_m) p(\mathbf{R}_t \mid \boldsymbol{\gamma}_j, \mathbf{Z}_{t-1}) \right], \tag{6}$$

$$\textit{where} \quad p(\mathbf{R}_t \mid \boldsymbol{\gamma}_j, \mathbf{Z}_{t-1}) = \frac{1}{(2\pi)^{N/2}} \sqrt{\frac{|\boldsymbol{\Lambda}_{0|\boldsymbol{\gamma}_j}|}{|\boldsymbol{\Lambda}_N|}} \frac{v_0^{S_0}}{v_N^{S_N}} \frac{\Gamma(S_N)}{\Gamma(S_0)}, \tag{7}$$

*$\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}dx$ is the gamma function, and $\pi(\boldsymbol{\gamma}_{j,\mathbf{f}} = I_m) = \prod_{k=1}^{K} w_k^{\gamma_{j,\mathbf{f},k}}(1-w_k)^{(1-\gamma_{j,\mathbf{f},k})}$ is the prior probability of the model as shown in (4). $I_m$ is a length $K$ vector with all elements 0 or 1, and $\{I_m\}_{m=1}^{2^K}$ enumerates all possible values that $\boldsymbol{\gamma}_{j,\mathbf{f}}$ can take. Note that $\boldsymbol{\gamma}_j$ is uniquely defined once $\boldsymbol{\gamma}_{j,\mathbf{f}}$ is given. Furthermore,*

$$\boldsymbol{\Lambda}_N = \mathcal{W}^\intercal \mathcal{W} + \boldsymbol{\Lambda}_{0|\boldsymbol{\gamma}_j}, \qquad \boldsymbol{\mu}_N = \boldsymbol{\Lambda}_N^{-1} \left( \mathcal{W}^\intercal \mathbf{r} \right),$$
$$v_N = v_0 + \frac{1}{2} \left( \mathbf{R}_t^\intercal \mathbf{R}_t - \boldsymbol{\mu}_N^\intercal \boldsymbol{\Lambda}_N \boldsymbol{\mu}_N \right), \qquad S_N = S_0 + \frac{N}{2},$$

*where $N$ is total number of data observations at time $t$, $\mathcal{W} = [\mathbf{1}, \mathbf{Z}_{t-1}]$, and $\boldsymbol{\Lambda}_{0|\boldsymbol{\gamma}_j}$ is a diagonal matrix with $s$-th element being $\xi_1^{-2}$ if $\gamma_{j,s} = 1$, or $\xi_0^{-2}$ if $\gamma_{j,s} = 0$ and 0 elsewhere.*

See proof in Internet Appendix I. To simplify the notation, we now denote the marginal likelihood evaluation of (6) using all observations in node $\mathcal{A}_n$, $n \in \{0, 1, 2, \dots\}$ by

$p(\mathcal{A}_n)$.

### 3.3.1 Post-Clustering Inference

Once the tree growth is complete, we refit the model using (2) for each cluster and obtain posterior samples of all model parameters using a Markov chain Monte Carlo (MCMC) approach, employing the Gibbs sampler for inference. Expressing the stacked data in matrix form for each time $t$, we have

$$\mathbf{R}_t = \mathbf{1}\mathrm{R}_{z,t} + \mathbf{Z}_{t-1}\mathbf{f}_t + \epsilon_{i,t} = \mathcal{W}_t\boldsymbol{\beta}_t + \boldsymbol{\epsilon}, \tag{8}$$

where $\mathcal{W}_t = [\mathbf{1}, \mathbf{Z}_{t-1}]$ and $\boldsymbol{\beta}_t = [\mathrm{R}_{z,t}, \mathbf{f}_t^{\mathsf{T}}]^{\mathsf{T}}$ to simplify notations. The full conditionals are given as follows,

1. For each month $t$, update $\boldsymbol{\beta}_t = \{\beta_k\}_{k=0,\cdots,K}$. Note that this step updates the regression coefficients of all variables. For $k = 0, 1, \cdots, K$,

$$\boldsymbol{\beta}_t \mid \sigma^2_{j,t}, \boldsymbol{\gamma}_j, \mathbf{R}_t, \mathcal{W}_t \sim \mathcal{N}\left(\tilde{\boldsymbol{\beta}}_t, \boldsymbol{V}\right),$$

   where $\tilde{\boldsymbol{\beta}}_t = (\mathcal{W}_t^{\mathsf{T}}\mathcal{W}_t + \boldsymbol{V}_0)^{-1}\mathcal{W}_t^{\mathsf{T}}\mathbf{R}_t$ and $\boldsymbol{V} = (\mathcal{W}_t^{\mathsf{T}}\mathcal{W}_t + \boldsymbol{V}_0)^{-1}\sigma^2_{j,t}$. $\boldsymbol{V}_0$ is a diagonal matrix where the $k$th diagonal elements is $\xi_{\gamma_{j,k}}$, updated each run after $\boldsymbol{\gamma}_{j,\mathbf{f},k}$ change.

2. Update $\boldsymbol{\gamma}_{j,\mathbf{f},k}$. Since the marginal likelihood is available in closed form (Proposition 1), we use a Collapsed Gibbs Sampler for $\boldsymbol{\gamma}$. This integrates out the coefficients $\boldsymbol{\beta}$ and variance $\boldsymbol{\sigma}^2$, improving mixing. For $k = 1, 2, \cdots, K$,

$$\gamma_{j,\mathbf{f},k} \mid \boldsymbol{\beta}, \sigma^2_j, \mathbf{R}, \mathcal{W} \sim \textsc{Ber}\left(\frac{p(\mathbf{R}|\gamma_{j,\mathbf{f},k}=1)w_k}{p(\mathbf{R}|\gamma_{j,\mathbf{f},k}=1)w_k + p(\mathbf{R}|\gamma_{j,\mathbf{f},k}=0)(1-w_k)}\right).$$

   where $p(\mathbf{R}|\gamma_{j,\mathbf{f},k}=1, \boldsymbol{\gamma}_{j,\mathbf{f},-k}, \mathbf{Z}) = \prod_{t=1}^{T} p(\mathbf{R}_t \mid \gamma_{j,\mathbf{f},k}=1, \boldsymbol{\gamma}_{j,\mathbf{f},-k}, \mathbf{Z}_t)$ is the marginal likelihood defined in equation (7), but with the $k$th variable selected. Similarly, $p(\mathbf{R}|\gamma_{j,\mathbf{f},k}=1, \boldsymbol{\gamma}_{j,\mathbf{f},-k}, \mathbf{Z})$ is the marginal likelihood with the $k$th varaible not selected.

3. Update $\sigma^2_{j,t}$, for $t = 1, 2, ..., T$

$$\sigma^2_{j,t} \mid \boldsymbol{\beta}_t, \boldsymbol{\gamma}_j, \mathbf{R}_t, \mathcal{W}_t \sim IG(S^*, v^*),$$

where $S^* = \frac{1}{2}(N + K + S_0)$, $v^* = \frac{1}{2}\left(||\mathbf{R} - \mathcal{W}\boldsymbol{\beta}||_2^2 + \sum_{k=0}^{K} \frac{\beta_k^2}{\xi_{\gamma_{j,k}}} + v_0\right)$, and $N$ is number of data observations in the leaf node at time $t$.

### 3.3.2 Testing for Heterogeneous Risk Premia

To conduct statistical inference on the factor risk premia based on the output from the Gibbs sampler, we employ a variance decomposition strategy based on Rubin's rules for multiple imputation (Rubin, 1987). This framework rigorously accounts for two distinct sources of uncertainty: the within-imputation time-series variation and the between-imputation parameter uncertainty captured by the posterior MCMC samples. Let $\boldsymbol{\lambda}$ denote the vector of risk premia (expected factor returns), defined as $\boldsymbol{\lambda} = \mathbb{E}[\mathbf{f}_t]$. We seek to estimate $\boldsymbol{\lambda}$ and its associated standard error using $S$ posterior draws obtained from the Gibbs sampler.

Let $\{\mathbf{f}_t^{(s)}\}_{t=1}^T$ denote the time-series of factors recovered in the $s$-th MCMC iteration, for $s = 1, \ldots, S$.

**Single-Group Rubin Combination** For each posterior draw $s$, we first compute the point estimate of the risk premia as the time-series average of the factors:

$$\hat{\boldsymbol{\lambda}}^{(s)} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{f}_t^{(s)}.$$

We estimate the associated variance of this mean for factor $k$, denoted $\mathrm{W}_k^{(s)}$:

$$\mathrm{W}_k^{(s)} = \frac{1}{T-1}\sum_{t=1}^{T} \left(\mathrm{f}_{t,k}^{(s)} - \hat{\lambda}_k^{(s)}\right)^2.$$

The final estimate of the risk premia $\bar{\boldsymbol{\lambda}}$ is the average over the $S$ posterior imputations:

$$\bar{\boldsymbol{\lambda}} = \frac{1}{S}\sum_{s=1}^{S} \hat{\boldsymbol{\lambda}}^{(s)}.$$

The total posterior covariance matrix $\mathbb{V}_{total}$ combines the within-imputation variance ($\bar{\mathrm{W}}_k$) and the between-imputation variance ($\mathbf{B}$):

$$\bar{\mathrm{W}}_k = \frac{1}{S}\sum_{s=1}^{S} \mathrm{W}_k^{(s)}, \quad \mathbf{B} = \frac{1}{S-1}\sum_{s=1}^{S}\left(\hat{\boldsymbol{\lambda}}^{(s)} - \bar{\boldsymbol{\lambda}}\right)\left(\hat{\boldsymbol{\lambda}}^{(s)} - \bar{\boldsymbol{\lambda}}\right)^{\top}.$$

Following Rubin (1987), the total variance is given by:

$$\mathbb{V}_{total} = \begin{bmatrix} \bar{\mathrm{W}}_1 & & \\ & \ddots & \\ & & \bar{\mathrm{W}}_k \end{bmatrix} + \left(1 + \frac{1}{S}\right)\mathbf{B}.$$

The scalar term $(1 + 1/S)$ adjusts for the finite number of posterior samples. The standard error for the $k$-th factor premium is the square root of the $k$-th diagonal element of $\mathbb{V}_{total}$.

**Testing for Heterogeneity** To test whether one factor in two clusters (Leaf $A$ and Leaf $B$) possesses statistically different risk premium, we construct the difference statistic $\boldsymbol{\delta} = \bar{\boldsymbol{\lambda}}_A - \bar{\boldsymbol{\lambda}}_B$. Under the assumption that the time-series errors are independent across clusters conditional on the factors, the standard error of the difference is:

$$\mathrm{SE}(\delta_k) = \sqrt{[\mathbb{V}_{total,A}]_{kk} + [\mathbb{V}_{total,B}]_{kk}}.$$

## 4 Uncommon Factors in the Cross Section of U.S. Equities

This section presents the empirical findings of applying the baseline BFCM to U.S. equity returns. Throughout the analysis, we use the agnostic investor prior $w_i = 0.5$, indicating an equal prior probability of selecting any factor.

### 4.1 Economically Guided Asset Clusters

BFCM clusters the unbalanced panel of individual stock returns by iteratively splitting the entire sample using firm characteristics, guided by marginal likelihood improvement as the global split criterion. Thus, the splitting variables that appear

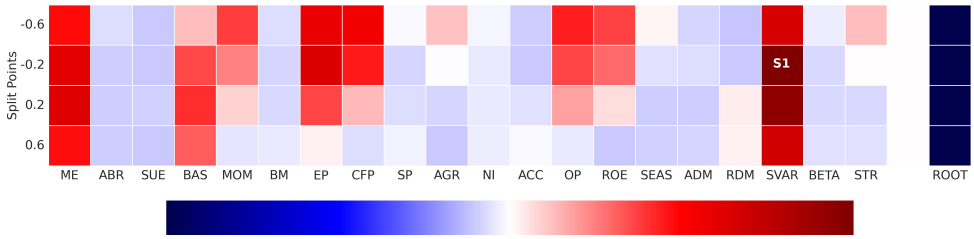
earlier are more important, as they contribute to higher marginal likelihood improvement, serving as evidence of clustering.

The heat map in Figure 2 illustrates the marginal likelihoods for all 80 potential split candidates, with the top value achieved by splitting on SVAR at a threshold of -0.2 (S1), followed by splits on SVAR at 0.6 (S2) and ME at 0.2 (S3). The prominence of volatility and size in partitioning the cross section of individual equity returns aligns with findings in the broader finance literature. Furthermore, all split candidates yield higher marginal likelihoods than the root, underscoring the presence of asset heterogeneity and the advantages of clustering.

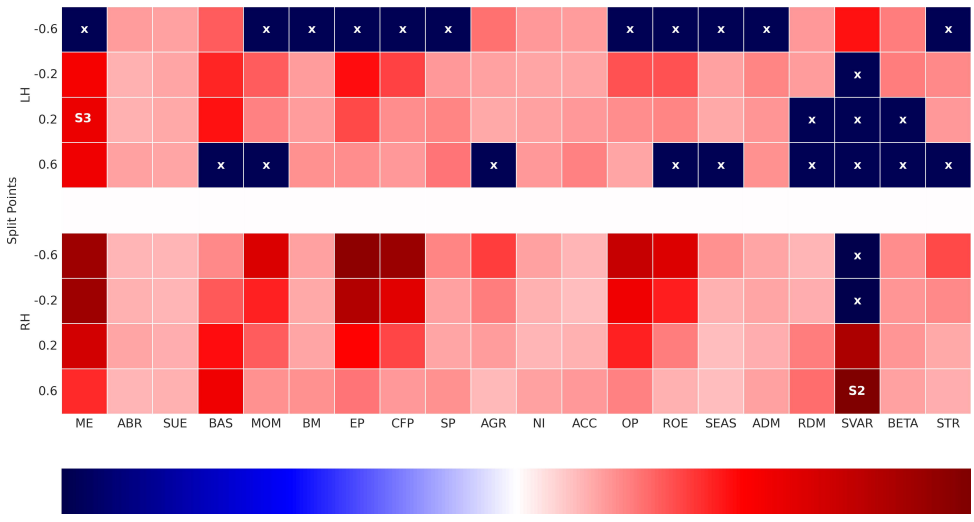
Figure 2: First Three Splits in the Cross Section

The heat map in panel (a) displays the search for the first split's cutpoint, aiming for the highest marginal likelihood value. Twenty characteristics and their cross-sectional quintiles are considered as candidates for cutpoints. The heat map in panel (b) shows the search for the second split's cutpoint, also aiming for the highest marginal likelihood value. The tree algorithm must compute 80 $(K + 1)$ candidates for iterative splitting.

(a) First Split




(b) Second and Third Split



The first split results in two leaf nodes, and the second split can occur on either side (see Figure A.3). The heat maps in the lower panels display the evaluations of split criteria for all candidate splits. Each panel represents potential splits at the left and right leaf nodes, resulting in $80 \times 2$ split candidates. After the first two splits based on variance (SVAR), the whole cross-section is partitioned into three parts. We explicitly examine the heterogeneity of risk premia within this economically guided partition before proceeding to the full tree structure, as the initial splits capture the dominant drivers of variation and demonstrate the strongest evidence of segmentation.

Table 3 reports estimates and the factor selection probabilities in Panel A, along with the $p$-values for the equality test in Panel B. The results show that the factor selection and risk premium estimates all exhibit group heterogeneity. For example, the High SVAR cluster doesn't select BM, while the other two clusters select with a high probability. However, the estimate of BM in Low SVAR is significantly different from the estimate overall. STR is selected by all clusters; yet, the estimates are significantly different from the overall. With the exception of the Low-Mid comparison, the estimates for the other two pairs differ significantly. In summary, the cluster found by BFCM exhibits stronger evidence of heterogeneity than the simple size group.

After the first two splits, there are still splitting candidates that can further increase the marginal likelihood. Allowing more splits not only enhances statistical significance (as indicated by the marginal likelihood increase) but also improves the model's predictive performance before it overfits.[4] For brevity, we omit demonstrations for subsequent splits. $K$ splits yield $K+1$ leaves, and the algorithm considers $80 \times (K+1)$ possible candidates for iterative splitting. Figure A.1 (in the appendix) displays the improvement in (log) marginal likelihood for each split compared to the root node, accounting for the normalizing constants. The initial split significantly increases the marginal likelihood, as do many subsequent splits, corroborating the presence of grouped heterogeneity. As the benefit of additional splits eventually diminishes, tree

[4]During certain periods, specific split candidates may yield clusters with insufficient observations due to sparse data on interacting characteristics. To address this, we impose a minimum leaf size of 50 stock returns per month to ensure a sufficient number of observations. Split candidates that fail to meet this criterion are discarded, as denoted by dark cells with a white cross (X) in Figure 2.

Table 3: **Test of Equality of Risk Premia after First Two Splits**

Panel A of this table presents the Bayesian estimate of factor risk premia for each cluster, computed from 2,000 MCMC samples, with the columns representing clusters after the first two splits (defined by SVAR) and overall pooled data. The selection probability (in percent) is also presented in parentheses. Panel B gives the $p$-value of the test of equality of risk premia between each group and the overall based on section 3.3.2. The sample period is based on the in-sample data (1980-2010).

| | Low SVAR | Mid SVAR | High SVAR | Overall | Low-Overall | Mid-Overall | High-Overall | Low-Mid | Low-High | Mid-High |
|---|---|---|---|---|---|---|---|---|---|---|
| | Panel A: Estimate and Selection Prob | | | | Panel B: $p$-value | | | | | |
| ME | -0.14(100) | -0.24(100) | -1.17(100) | -0.23(100) | 0.371 | 0.894 | 0.000 | 0.362 | 0.000 | 0.000 |
| ABR | 0.01(0) | 0.04(0) | 0.05(0) | 0.09(0) | 0.000 | 0.008 | 0.041 | 0.013 | 0.065 | 0.966 |
| SUE | 0.61(100) | 1.01(100) | 1.39(100) | 0.96(100) | 0.000 | 0.392 | 0.000 | 0.000 | 0.000 | 0.000 |
| BAS | -0.00(0) | -0.01(0) | -0.03(0) | -0.03(0) | 0.071 | 0.141 | 0.871 | 0.953 | 0.115 | 0.169 |
| MOM | 0.05(100) | 0.20(100) | 0.01(0) | 0.15(100) | 0.298 | 0.650 | 0.054 | 0.143 | 0.548 | 0.017 |
| BM | 0.13(100) | 0.27(100) | 0.07(0) | 0.32(100) | 0.005 | 0.477 | 0.000 | 0.097 | 0.308 | 0.006 |
| EP | 0.17(100) | 0.02(0) | 0.03(0) | 0.18(100) | 0.882 | 0.001 | 0.002 | 0.019 | 0.027 | 0.788 |
| CFP | 0.19(100) | 0.02(0) | 0.05(0) | 0.02(0) | 0.006 | 0.740 | 0.156 | 0.008 | 0.027 | 0.243 |
| SP | 0.04(100) | 0.21(100) | 0.06(0) | 0.22(100) | 0.024 | 0.929 | 0.009 | 0.059 | 0.739 | 0.042 |
| AGR | -0.01(0) | -0.03(0) | -0.04(0) | -0.06(0) | 0.000 | 0.062 | 0.306 | 0.168 | 0.117 | 0.624 |
| NI | -0.01(0) | -0.03(0) | -0.05(0) | -0.05(0) | 0.004 | 0.312 | 0.954 | 0.094 | 0.027 | 0.382 |
| ACC | -0.03(0) | -0.04(0) | -0.03(0) | -0.07(0) | 0.001 | 0.065 | 0.024 | 0.229 | 0.997 | 0.432 |
| OP | 0.01(0) | 0.03(0) | 0.03(0) | 0.05(0) | 0.005 | 0.222 | 0.286 | 0.195 | 0.358 | 0.953 |
| ROE | 0.01(0) | 0.03(0) | 0.02(0) | 0.04(0) | 0.011 | 0.281 | 0.251 | 0.219 | 0.552 | 0.779 |
| SEAS | 0.01(0) | 0.01(0) | 0.02(0) | 0.03(0) | 0.155 | 0.300 | 0.784 | 0.835 | 0.401 | 0.545 |
| ADM | -0.00(0) | 0.00(0) | 0.02(0) | 0.00(0) | 0.911 | 0.989 | 0.467 | 0.922 | 0.373 | 0.457 |
| RDM | 0.01(0) | 0.06(0) | 0.06(0) | 0.10(0) | 0.000 | 0.021 | 0.088 | 0.001 | 0.007 | 0.849 |
| SVAR | 0.09(100) | -0.30(100) | -1.25(100) | -0.21(100) | 0.023 | 0.544 | 0.003 | 0.016 | 0.000 | 0.009 |
| BETA | 0.10(100) | 0.12(100) | 0.00(0) | 0.09(100) | 0.875 | 0.795 | 0.165 | 0.919 | 0.145 | 0.125 |
| STR | -0.44(100) | -0.51(100) | -1.16(100) | -0.70(100) | 0.001 | 0.031 | 0.000 | 0.388 | 0.000 | 0.000 |

growth stops after 8 splits.

The BFCM tree (Figure 3) has 9 terminal leaves, representing 9 clusters managed through multiple equity characteristics. The sequence of split rules from the root to the terminal leaves, based on characteristics, determines an asset's cluster membership. This membership may change as asset characteristics change. The average number of monthly return observations for each terminal leaf node is reported, showing that the resulting leaf clusters are well-balanced, neither too large nor too small.

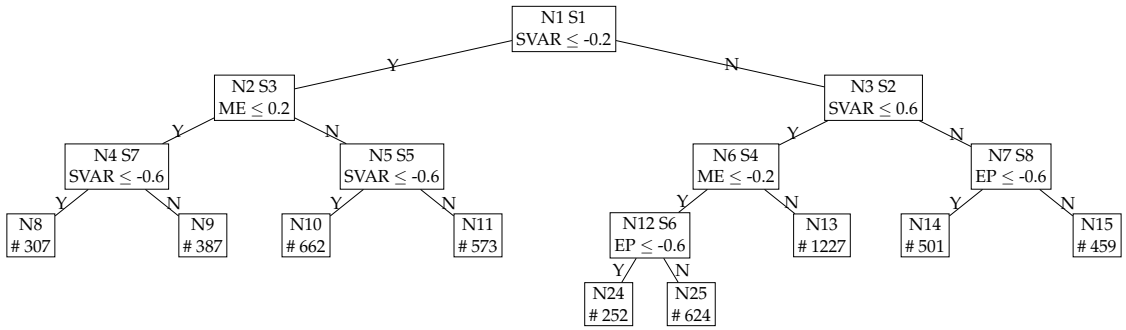
In the third and fourth layers, EP serves as the third important splitting variable in addition to SVAR and ME, while other characteristics show a limited effect. This characteristic-based clustering of the cross section captures asset heterogeneity, with clusters potentially exhibiting different factor structures, and are used as test assets on the left-hand side to form CS factors in Equation (1).

### 4.2 Common and Uncommon CS Factors in the Cross Section

**Factor selection and common factors.** Table 4 presents the Bayesian estimates of the risk premia for various factors, computed from 2,000 MCMC samples. The IDs of leaf

Figure 3: Tree Cluster

Using monthly data from 1980 to 2010, the tree divides individual stock returns based on cross-sectional standardization of firm characteristic ranks to [-1,1]. The resulting terminal leaves correspond to clusters identified by the firm characteristics. Each node, including leaves and intermediate nodes, has an ID indicated by $N\#$, and the order of the splits is denoted by $S\#$. For instance, $S3$ refers to the third split. The number in the bottom nodes is the average number of stocks per month.




clusters correspond to the terminal leaves of the tree in Figure 3. In parentheses, we report the factor selection probabilities in each leaf cluster, which is the posterior mean of $\gamma_{j,\mathbf{f}}$. Note that we estimate the CS factor for each time period and take the average across time to obtain the risk premia estimate. The intercept of each cross-sectional regression, by definition, represents a "level return" that represents the "market" in the context of CS factors (Fama and French, 2020). Thus, we give a slab prior for the intercept, meaning that the market is always the common factor.

The extreme evidence of posterior selection probabilities to the boundaries of 0 or 1 is a structural implication of applying spike-and-slab priors to the factor realizations at each time $t$. Because the variable selection mechanism evaluates the existence of a risk premium on a period-by-period basis, the resulting probability represents the accumulation of pricing evidence across the entire time series.

ME is selected with extremely high probability in most leaf clusters, the sole exception being N24 (0%). This finding confirms the market equity or size's role as a universal pricing factor, aligning with expectations from Fama and French (1993). SUE and SVAR are similarly prominent, with selection probabilities greater than 0.95 in 5

Table 4: **Factor Selection Probability and Risk Premia Estimates**

This table presents the Bayesian estimate of factor risk premia for each leaf cluster, computed from 2,000 MCMC samples, with the leaf cluster IDs matching those of the tree in Figure 3. The selection probability (in percent) is also presented in parentheses. The sample period is the in-sample data (1980-2010).

| Factor | N8 | N9 | N10 | N11 | N24 | N25 | N13 | N14 | N15 |
|---|---|---|---|---|---|---|---|---|---|
| ME | 0.09(100) | -0.08(100) | -0.25(100) | -0.24(100) | -0.01(0) | -0.95(100) | -0.28(100) | -1.88(100) | -1.13(100) |
| ABR | 0.00(0) | 0.00(0) | 0.00(0) | 0.01(0) | 0.00(0) | 0.02(0) | 0.03(0) | 0.03(0) | 0.02(0) |
| SUE | 0.86(100) | 1.00(100) | 0.02(0) | 0.02(0) | 0.02(0) | 1.57(100) | 0.67(100) | 0.05(0) | 1.38(100) |
| BAS | -0.01(0) | -0.00(0) | -0.00(0) | -0.00(0) | 0.00(0) | -0.00(0) | -0.00(0) | -0.02(0) | -0.02(0) |
| MOM | 0.00(0) | 0.00(0) | 0.11(100) | 0.31(100) | 0.00(0) | 0.01(0) | 0.28(100) | 0.01(0) | 0.01(0) |
| BM | 0.01(0) | 0.01(0) | -0.05(100) | 0.07(100) | 0.01(0) | 0.02(0) | 0.01(0) | 0.04(0) | 0.04(0) |
| EP | 0.01(0) | 0.01(0) | 0.24(100) | 0.01(0) | -0.00(0) | 0.02(0) | 0.15(100) | -1.26(100) | 0.02(0) |
| CFP | 0.01(0) | 0.01(0) | 0.05(100) | 0.01(0) | 0.00(0) | 0.02(0) | 0.01(0) | 0.01(0) | 0.03(0) |
| SP | 0.00(0) | 0.01(0) | 0.04(100) | 0.13(100) | 0.01(0) | 0.02(0) | 0.28(100) | 0.03(0) | 0.03(0) |
| AGR | -0.00(0) | -0.00(0) | -0.00(0) | -0.00(0) | -0.01(0) | -0.01(0) | -0.02(0) | -0.03(0) | -0.01(0) |
| NI | -0.00(0) | -0.00(0) | -0.00(0) | -0.01(0) | -0.01(0) | -0.01(0) | -0.02(0) | -0.02(0) | -0.02(0) |
| ACC | -0.01(0) | -0.01(0) | -0.01(0) | -0.01(0) | -0.00(0) | -0.02(0) | -0.03(0) | -0.00(0) | -0.02(0) |
| OP | 0.01(0) | 0.01(0) | 0.00(0) | 0.01(0) | 0.00(0) | 0.01(0) | 0.01(0) | 0.01(0) | 0.01(0) |
| ROE | 0.01(0) | 0.01(0) | 0.00(0) | 0.01(0) | 0.01(0) | 0.01(0) | 0.01(0) | 0.01(0) | 0.01(0) |
| SEAS | 0.00(0) | 0.00(0) | 0.00(0) | 0.00(0) | 0.01(0) | 0.01(0) | 0.00(0) | 0.01(0) | 0.01(0) |
| ADM | 0.00(0) | 0.00(0) | 0.00(0) | -0.00(0) | 0.01(0) | 0.01(0) | 0.00(0) | 0.01(0) | 0.01(0) |
| RDM | 0.00(0) | 0.00(0) | 0.00(0) | 0.00(0) | 0.02(0) | 0.02(0) | 0.03(0) | 0.05(0) | 0.02(0) |
| SVAR | -0.00(0) | -0.00(0) | -0.26(100) | -0.00(0) | 0.00(0) | -0.04(100) | -0.47(100) | -2.06(100) | -0.85(100) |
| BETA | 0.00(0) | 0.00(0) | 0.15(100) | 0.04(100) | 0.00(0) | 0.01(0) | 0.09(100) | 0.00(0) | 0.00(0) |
| STR | 0.00(0) | -0.01(0) | -0.47(100) | -0.50(100) | -0.01(0) | -0.71(100) | -0.39(100) | -0.08(0) | -0.06(0) |

out of 9 clusters. Their usefulness overlaps several clusters (N25, N13, N15), but differs in others. However, the remaining factors demonstrate limited usefulness, as they appear in only a few clusters. For instance, the cash flow factors (CFP) are selected with probabilities above 0.95 in just 1 cluster.

N24 is a unique cluster that excludes any factors. It represents assets with median volatility ($-0.2 <$ SVAR $\leq 0.6$), poor earnings (EP $\leq -0.2$), low earnings (EP $\leq -0.6$) and small size (ME $\leq -0.2$). These characteristics are typically associated with the most challenging stocks in empirical asset pricing (see, e.g., Fama and French, 1996, 2015). In comparison, N25, which shares the same parent node as N24, differs slightly in EP as a splitting rule but has a high selection probability for the factor of ME, SUE, SVAR, and STR. This provides supporting evidence for asset heterogeneity, with BFCM identifying that N24 assets are the least "priceable."

**Uncommon factors.** While ME is the most common factor, many other factors are selected with high probability only in a subset of clusters, and are, therefore, uncommon

and asset-specific. To understand which factors price what assets, we further summarize the factor selection results in Table 4 using a probability threshold in Table 5. Specifically, we consider the threshold of 0.95 to differentiate factors based on their selection probabilities. This way, we clarify which factors are common across all clusters and which are uncommon and specific to different clusters.

Table 5: Factor Selection Evaluation by Leaf Clusters

This summary table presents the results from Table 4 under 95% selection probability thresholds. Four panels display leaf cluster positions for the top three tree layers. The selection is based on the in-sample period (1980-2010).

| Prob. | N8 | N9 | N10 | N11 | N24 | N25 | N13 | N14 | N15 |
|---|---|---|---|---|---|---|---|---|---|
| | Panel A: I(SVAR $\leq$ -0.6 & ME $\leq$ 0.2) | | Panel B: I(SVAR $\leq$ -0.6 & ME >0.2) | | Panel C: I(-0.2 $<$ SVAR $\leq$ 0.6) | | | Panel D: I(SVAR >0.6) | |
| $>$ 0.95 | ME<br>SUE | ME<br>SUE | ME<br>MOM<br>BM<br>EP<br>CFP<br>SP<br>SVAR<br>BETA<br>STR | ME<br>MOM<br>BM<br>SP<br>BETA<br>STR | | ME<br>SUE<br>SVAR<br>STR | ME<br>SUE<br>MOM<br>EP<br>SP<br>SVAR<br>BETA<br>STR | ME<br>EP<br>SVAR | ME<br>SUE<br>SVAR |

Table 5 summarizes the factor selection patterns in the mega clusters based on the splits in the top two layers. Panel A (i.e., characteristics-managed asset clusters N8 and N9) includes stocks with the lowest volatility and low size (i.e., SVAR $\leq -0.6$ & ME $\leq 0.2$). We find that SUE is crucial for these stocks beyond ME, as they are selected with a probability greater than 0.95. Panel B includes clusters of low variance stocks and large size (i.e., SVAR $\leq -0.6$ & ME $> 0.2$). ME, MOM, BM, SP, BETA, and STR are the most important factors for these sub-clusters, but N10 requires additional EP, CFP, and SVAR.

Panel C consists of 8 clusters with relatively high variance stocks ($-0.2 <$ SVAR $\leq 0.6$). These clusters exhibit significant heterogeneity, requiring distinct factors, with the extreme case of no factor in N24. Panel D includes two clusters with the highest volatility stocks (SVAR $> 0.6$) and involves only three factors for each. Remarkably, SVAR is an important factor in these two clusters. Many other factors lack power in this extreme case.

### 4.3 Predictive Performance

An asset pricing model must be evaluated in terms of its empirical performance. Typically, performance is evaluated through cross-sectional $R^2$ ($CSR^2$). However, the $CSR^2$ metric can be misleading (Lewellen et al., 2010), as it reflects only in-sample fit, and a large increase in $CSR^2$ may not be statistically significant (Kan et al., 2013).

In the CS factor approach, the average of the slope from the period-by-period regression can be interpreted as a predictive signal. BFCM enables the fitting of individual stocks in different clusters separately to form cluster-specific CS factors, thereby facilitating heterogeneous factor selection and estimation. While BFCM relies on economically guided statistical fitness (marginal likelihood of factor models), its predictive power, especially out-of-sample, is not guaranteed and depends on the validity of the assumed grouped heterogeneity with uncommon factors.

To evaluate the predictive performance in a specific cluster ($j$), we rely on the Predictive $R^2$ following Gu et al. (2020), defined as

$$\text{Predictive } R^2 = 1 - \frac{\sum_{i=1}^{N_j}\sum_{t=1}^{T_i}\left(r_{i,t} - \widehat{r}_{i,t}\right)^2}{\sum_{i=1}^{N_j}\sum_{t=1}^{T_i} r_{i,t}^2}, \tag{9}$$

where $\widehat{r}_{i,t}^{(j)} = \sum_{j=1}^{J} \mathbb{1}_{\{\widehat{T}(\mathbf{z}_{i,t-1})=j\}}(\mathbf{z}_{j,t}^{\mathsf{T}}\hat{\mathbf{f}}_{j,t} + \hat{r}_{j,z,t})$, and $N_j$ is number of assets in the $j$-th cluster. It measures how well the model predicts returns across all individual stocks in cluster $j$. We compare the performance of the heterogeneous factor model with the OLS CS factor model (without factor selection) in each of the 9 clusters. A higher out-of-sample Predictive $R^2$ value suggests better model predictive power, and a positive value confirms that the factor model has better fitness than the naive zero benchmark. We use the OLS CS factor model as our benchmark because it is arguably the standard way and is widely adopted.

Table 6 shows that the uncommon-factor models outperform the cluster-specific OLS and pooled models in the clusters when stock variance is low (Panels A and B). The small variance stocks tend to be more predictable, with a much higher predictive $R^2$ than large variance stocks. In contrast, Panels C and D show that for clusters with

higher stock variance, the predictive power of cluster-specific factors diminishes, often falling below that of a global factor model. However, even the global factor model exhibits limited predictive power, with $R^2$ values close to zero in these cases. These results emphasize the challenges of prediction in high-variance stock clusters while underscoring the relative advantages of uncommon-factor models in low-variance scenarios.

The BFCM improvement in predictive performance comes from both Bayesian factor selection and characteristic-based clustering. The latter captures grouped heterogeneity, while the former reflects the implications of this heterogeneity. Identifying the heterogeneity group is extremely useful because BFCM informs us about the predictive power of different assets. Some clusters can be understood by a model with selected factors, while others are very noisy.

Table 6: Evaluating Predictive Performance with Leaf Clusters

This table presents the out-of-sample predictive $R^2$ (%) values for different models, computed for each leaf cluster to evaluate pricing errors in fitted factor models. We include Fama and French (2020) OLS CS factors in both cluster-specific and pooled samples. The leaf node IDs match the tree cluster in Figure 3.

| Prob. | N32 | N33 | N10 | N22 | N24 | N25 | N13 | N14 | N15 |
|---|---|---|---|---|---|---|---|---|---|
| | Panel A: I(SVAR $\leq$ -0.6 & ME $\leq$ 0.2) | | Panel B: I(SVAR $\leq$ -0.6 & ME $>$0.2) | | Panel C: I(-0.2 $<$ SVAR $\leq$ 0.6) | | | Panel D: I(SVAR $>$0.6) | |
| Bayes | 1.24 | 1.73 | 1.67 | 0.83 | 0.08 | 1.1 | 0.35 | -0.04 | 0.07 |
| OLS | 0.86 | 1.63 | 1.65 | 0.78 | 0.24 | 1.12 | 0.26 | -0.18 | -0.04 |
| Bayes Pool | 0.88 | 1.57 | 0.05 | -0.16 | 0.25 | 1.05 | 0.08 | 0.09 | 0.2 |
| OLS Pool | 0.73 | 1.55 | 0.07 | -0.24 | 0.26 | 1.04 | -0.07 | 0.07 | 0.16 |

### 4.4 Model-Predicted Investment Performance

In the last section, we assess the predictive performance based on clusters. Of particular interest is how cluster-specific predictions contribute to improving practical investment performance. In this section, we investigate the model-predicted long-short (LS) portfolio.

For the cluster portfolio, within each leaf cluster, we use the in-sample estimates of the slope coefficients obtained through Bayesian estimation and OLS Fama-MacBeth regression. The pooled estimation, in contrast, refers to estimating the slope coef-

ficients based on all stocks, without considering any clustering. During the out-of-sample period, at time $t$, we use the stock characteristics and the slope estimates to make return predictions for time $t+1$. Based on these predicted returns, we construct portfolios within each leaf cluster by going long on the top 20% of stocks and short on the bottom 20%, considering both equal-weighted and value-weighted portfolios. For the in-sample period, we can also construct a long-short portfolio, which is used to calculate the cluster tangency portfolio weight. Finally, we apply the cluster tangency portfolio weight to the out-of-sample returns across all leaf clusters to form a single investable portfolio. For the overall portfolio, we directly use the model's predictions to construct long-short (20%) portfolios based on all stocks, without differentiating by cluster.

Table 7 shows the summary statistics for the cluster and overall LS portfolios, both equal-weighted and value-weighted. The clustered portfolio with heterogeneous risk premia achieves the best performance in both the equal-weighted (first column of Panel A) and value-weighted (third column of Panel A) cases. The value-weighted portfolios experience a much sharper deterioration with pool coefficients, with Sharpe ratios falling below 1.

To verify whether the superior performance of our cluster-based strategies is truly novel or merely a compensation for known risk factors, we conduct spanning regressions. Since the equal-weighted portfolio is biased due to the inclusion of small-cap stocks, we focus on the value-weighted case (Bayes Coef Cluster VW). Table 8 reports the results of these spanning regressions. We consider six benchmark models, with factors including the FF5 factor, momentum factor (UMD), liquidity factor (LIQ) from Pástor and Stambaugh (2003), q-factor from Hou et al. (2015), quality minus junk (QMJ) from Asness et al. (2019), and betting-against-beta (BAB) from Frazzini and Pedersen (2014).

The results indicate that the alpha generated by our cluster-based portfolio remains economically large and statistically significant across all specifications. As shown in column 1, the portfolio generates an annualized alpha of approximately 2.11% ($t$-

stat = 12.42) relative to the Fama-French 5-factor model plus Momentum. Furthermore, the loadings on standard risk factors provide insight into the unique nature of our strategy. While there is some significant exposure to HML and occasional loading on BAB or UMD, these traditional factors fail to explain most of the returns. The high unexplained alpha suggests that the BFCM captures cluster-specific pricing information—"uncommon factors"—that is aggregated away or overlooked by common factor models.

Table 7: Model Predicted Long Short Portfolios

This table presents the long-short portfolios based on model prediction. The cluster portfolio forms long-short (20%) portfolios in each cluster and uses the in-sample (1980-2010) estimate to calculate the tangency portfolio weight, then applies it to out-of-sample (2011-2024) cluster portfolios. The overall portfolio forms long-short portfolios based on the full cross-section. We consider both equal-weighted (EW) and value-weighted (VW). Panel A uses cluster-wise coefficient estimates, while Panel B uses overall pool estimates.

| | Cluster EW | Overall EW | Cluster VW | Overall VW |
|---|---|---|---|---|
| | Panel A: Bayes Coef | | | |
| Mean Return | 2.37 | 2.04 | 2.08 | 0.49 |
| $t$-stat | (17.17) | (9.23) | (14.61) | (1.32) |
| Median Return | 2.20 | 2.08 | 1.83 | 0.52 |
| SD | 1.94 | 2.75 | 2.10 | 4.21 |
| Skewness | 0.07 | -0.15 | 0.35 | 0.02 |
| Kurtosis | 0.17 | 0.54 | 0.68 | 3.45 |
| AC1 | -0.14 | 0.02 | -0.20 | 0.06 |
| Sharpe Ratio | 4.23 | 2.57 | 3.44 | 0.40 |
| Max Drawdown | 0.04 | 0.07 | 0.04 | 0.39 |
| | Panel B: Bayes Pool Coef | | | |
| Mean Return | 2.62 | 2.17 | 2.28 | 0.41 |
| $t$-stat | (14.93) | (9.67) | (12.31) | (1.61) |
| Median Return | 2.45 | 2.31 | 2.12 | 0.56 |
| SD | 2.79 | 2.89 | 2.74 | 2.80 |
| Skewness | 0.38 | 0.12 | 0.19 | 0.34 |
| Kurtosis | 1.61 | 0.46 | 1.04 | 2.26 |
| AC1 | -0.19 | -0.03 | -0.21 | 0.09 |
| Sharpe Ratio | 3.26 | 2.61 | 2.89 | 0.51 |
| Max Drawdown | 0.11 | 0.11 | 0.06 | 0.30 |

### 4.5 Cluster-wise CS Factors

In addition to the long-short sorted portfolio based on model predictions discussed in the previous section, we also examine investing directly in the CS factor tangency portfolio, both with and without clustering.

Table 8: Spanning Regression of Cluster Long-short Portfolio

This table presents the spanning regression of Bayes Coef Cluster VW portfolio in Table 7 on several factor models. The factors include FF5 factor,momentum factor (UMD), liquidity factor (LIQ) from Pástor and Stambaugh (2003), q-factor from Hou et al. (2015), quality minus junk (QMJ) from Asness et al. (2019), betting-against-beta (BAB) from Frazzini and Pedersen (2014).

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Alpha | 2.113*** (12.422) | 2.069*** (12.149) | 2.067*** (12.339) | 2.100*** (11.991) | 2.035*** (11.524) | 2.071*** (11.424) |
| MktRf | 0.007 (0.179) | 0.028 (0.660) | 0.023 (0.602) | -0.003 (-0.061) | 0.009 (0.211) | -0.006 (-0.105) |
| SMB | 0.001 (0.014) | 0.029 (0.387) | | | | -0.238 (-0.833) |
| HML | 0.115 (1.609) | 0.154** (2.093) | | | | 0.130 (1.538) |
| RMW | -0.088 (-0.950) | -0.064 (-0.696) | | | | -0.017 (-0.125) |
| CMA | -0.169 (-1.585) | -0.209* (-1.939) | | | | -0.275 (-1.207) |
| UMD | | 0.102* (1.929) | | | | 0.070 (1.089) |
| LIQ | | | 0.057 (1.165) | | | 0.034 (0.613) |
| ME | | | | 0.078 (1.023) | | 0.249 (0.885) |
| IA | | | | -0.079 (-0.960) | | 0.059 (0.294) |
| ROE | | | | 0.048 (0.549) | | -0.036 (-0.260) |
| EG | | | | -0.073 (-0.668) | | -0.035 (-0.266) |
| QMJ | | | | | -0.090 (-1.248) | -0.036 (-0.274) |
| BAB | | | | | 0.124* (1.773) | 0.091 (1.139) |

The CS factor weight for the next time period is $(\mathbf{Z}_t^\intercal \mathbf{Z}_t)^{-1}\mathbf{Z}_t^\intercal$. To make sure the CS factor weight is in the same unit, we then rescale the weights so that the sums of the weights for the long and short legs are 1 and -1, respectively. Next, in each leaf cluster, we calculate the in-sample CS factors and use the mean and covariance to derive the tangency portfolio weights. We only consider the CS factors selected in the table 5. These tangency weights are then applied to the out-of-sample cluster CS factors. Note that these weights ($K \times 1$) correspond to each CS factor within a single cluster. We get one tangency CS factor portfolio for each cluster. We then use the in-sample cluster tangency portfolio's mean and covariance to construct a tangency portfolio (weights $J \times 1$, where $J$ is the number of clusters) for each cluster portfolio. For the overall CS factor portfolio, we simply repeat the cluster-specific operation for all stocks.

Table 9 presents the summary statistics and the spanning regression for the cluster and overall CS factors. The overall CS factor already has a good performance, with a Sharpe ratio of up to 3.21. However, considering the cluster will further improve the performance to an annualized Sharpe ratio of 4.20, with a much higher expected return of 3.03. Importantly, the spanning test reveals that after adjusting for Market, the overall factors are fully explained by the cluster CS factor, with an insignificant

alpha. In contrast, the cluster CS factor exhibits a highly significant unexplained alpha of 1.54.

Table 10 shows the performance of the CS factor portfolios in each cluster, along with the tangency portfolio weights. It indicates that not all cluster portfolios perform equally well; however, the in-sample tangency portfolios' weights are informative, suggesting that the large positive weights are assigned to the better-performing cluster out of sample. It should be noted that the slope factors constructed directly from individual stock regressions cannot be value-weighted. Consequently, the reported performance may be disproportionately driven by small-cap stocks. Therefore, we provide these results primarily for reference and may be less practical for implementation than the value-weighted long-short portfolio in section 4.4.

Table 9: Cluster-wise and Overall CS factor Portfolios

This table presents the summary statistics of the performance of the out-of-sample (2011-2024) cluster and overall CS factor portfolios, and the spanning regression between cluster and overall.

| Summary Stat | | | Spanning Regression | | | |
|---|---|---|---|---|---|---|
| | Cluster | Overall | | Slope | SE | $t$-stat |
| Mean Return | 3.03 | 1.18 | | Overall $\sim$ MktRf + Cluster | | |
| $t$-stat | (18.21) | (11.98) | (Intercept) | 0.13 | 0.12 | 1.10 |
| Median Return | 2.99 | 1.15 | MktRf | 0.03. | 0.02 | 1.85 |
| SD | 2.49 | 1.28 | Cluster | 0.34*** | 0.03 | 11.30 |
| Skewness | 0.15 | 0.43 | | | | |
| Kurtosis | 0.63 | 0.94 | | Cluster $\sim$ MktRf + Overall | | |
| AC1 | -0.17 | -0.08 | (Intercept) | 1.54*** | 0.20 | 7.72 |
| Sharpe Ratio | 4.20 | 3.21 | MktRf | -0.04 | 0.03 | -1.31 |
| Max Drawdown | 0.06 | 0.05 | Overall | 1.30*** | 0.11 | 11.30 |

## 5 Uncommon Factors with Regime Switching

Having established the existence of uncommon factors across characteristic-sorted clusters in the cross-section, we now turn to the time-series dimension. The macroeconomic environment is inherently dynamic, and investors continually update their beliefs and portfolio decisions as new information becomes available. Consequently, the factor risk premia are unlikely to remain stable over time. A large body of finance literature documents pervasive model instability and structural breaks in asset

Table 10: Individual Cluster CS factor Portfolios

This table presents the out-of-sample performance of each individual cluster's CS factor portfolio as well as the tangency portfolio weight estimated from in-sample data.

| | N8 | N9 | N10 | N11 | N25 | N13 | N14 | N15 |
|---|---|---|---|---|---|---|---|---|
| Mean Return | 1.12 | 1.91 | 0.16 | -1.71 | 7.75 | 1.25 | -0.62 | 4.51 |
| $t$-stat | (10.46) | (11.70) | (1.38) | (-2.35) | (17.85) | (7.34) | (-2.94) | (8.00) |
| Median Return | 1.00 | 1.68 | 0.21 | -1.13 | 7.38 | 1.34 | -0.69 | 4.62 |
| SD | 1.43 | 2.15 | 1.49 | 11.13 | 6.80 | 2.15 | 3.17 | 8.46 |
| Skewness | 0.24 | 0.57 | -0.52 | -0.30 | 0.21 | -0.02 | 0.10 | 0.02 |
| Kurtosis | 0.22 | 1.58 | 2.72 | 1.87 | 0.54 | -0.14 | 1.66 | 2.58 |
| AC1 | -0.02 | 0.00 | -0.14 | -0.20 | -0.27 | -0.03 | -0.18 | -0.15 |
| Sharpe Ratio | 2.70 | 3.07 | 0.36 | -0.53 | 3.95 | 2.02 | -0.68 | 1.84 |
| Max Drawdown | 0.06 | 0.04 | 0.15 | 0.67 | 0.13 | 0.11 | 0.27 | 0.26 |
| Tangency Weight | 0.44 | 0.16 | 0.4 | -0.01 | 0.21 | -0.13 | -0.2 | 0.12 |

pricing relationships. For example, Smith and Timmermann (2021, 2022) study the implications of structural breaks and provide evidence of abrupt changes in factor risk premia.

As described in Section 3, BFCM can be easily extended to account for interpretable time series regime-switching by splitting based on macroeconomic variables—such as market volatility, inflation, or the Treasury bill rate—that capture time-series heterogeneity in asset returns. This feature allows BFCM to uncover the economic forces driving returns across distinct, macro-instrumented asset-pricing regimes and provides a coherent interpretation of regime-dependent factor structures.

### 5.1 Macro Regimes in the CS Factor Model

To ensure the resulting model remains interpretable and parsimonious, we impose constraints on the tree's growth in the time-series dimension. In the regime-switching literature, the number of regimes is typically small, ranging from two to four. Thus, we restrict the maximum number of time-series splits to two, resulting in a maximum of three distinct economic regimes. This constraint prevents the model from identifying spurious, short-lived "regimes" driven by noise and ensures that the identified states correspond to meaningful economic phases. By jointly considering time-series and cross-sectional split candidates, BFCM endogenously selects the optimal split type at each node, thereby capturing both cross-sectional and time-series

heterogeneity in asset returns.

Figure 4 illustrates the resulting tree structure and the macroeconomic states endogenously selected by the model. Unlike standard recession indicators, the BFCM selects Aggregate Market liquidity (LIQ) as the primary splitting variable to maximize the marginal likelihood. The first node isolates periods where aggregate liquidity is below its long-term trend, defining a "Illiquid Market" regime (Regime 1). The model then further splits the "High liquidity" node based on Net Equity Issuance (NI). This secondary split distinguishes between periods where external financing is effectively frozen (Regime 2: Liquid and Low Issuance) and periods where firms actively raise capital(Regime 3: Liquid and High Issuance).

After the first two splits based on macroeconomic variables, each regime can be further splits based on firm characteristics in the cross section. Different from the pure cross section in section 4, advertisement-to-market (ADM) emerges as the most useful variable once time-series heterogeneity is accounted for. Beyond SVAR, AGR and RDM can be additional important characteristics for cross-sectional heterogeneity. SVAR only emerges once during the illiquid regime.

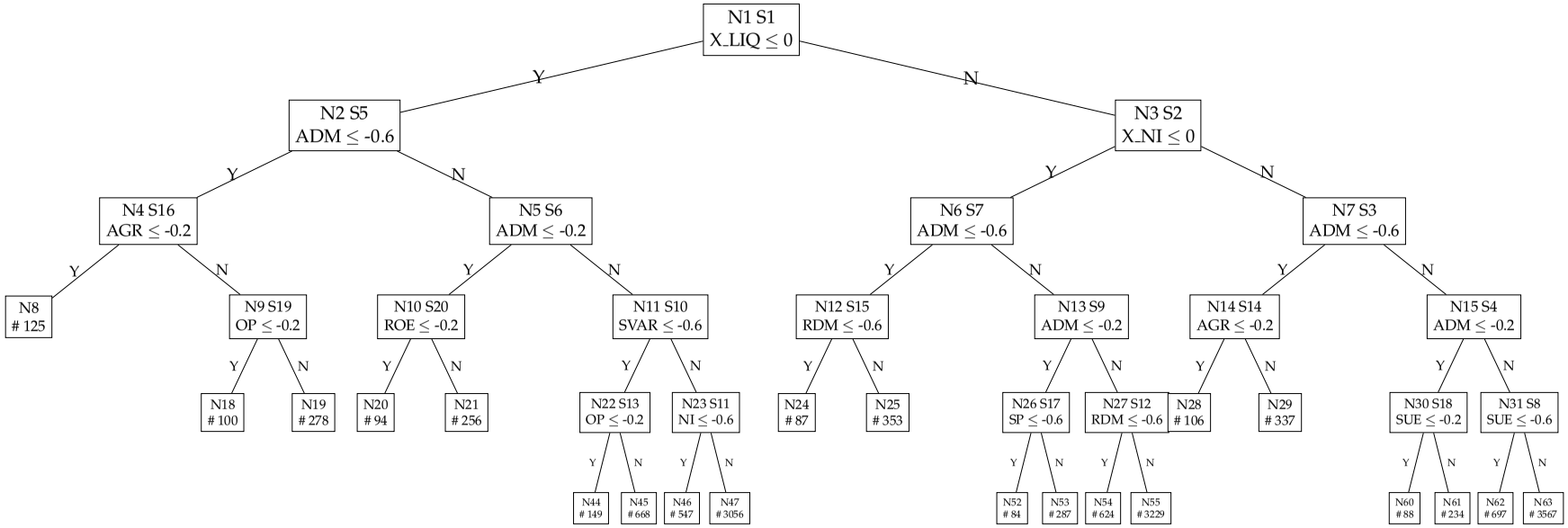
### 5.2 Uncommon CS Factors in both Time Series and Cross Section

The "uncommon factor" phenomenon observed in the cross-section is equally prevalent across these time-series regimes. Table 11 reports the estimated risk premia and factor selection probabilities and Table 12 summarizes the selected factors, revealing striking heterogeneity driven by economic regimes.

**Common Factors** While the BFCM identifies distinct regimes, certain systematic risks remain relevant across the time-series dimension. Stock variance (SVAR) remains a common factor selected in 13 out of 21 clusters, with significant representation across clusters within each regime. While the significance of market equity (ME) diminishes relative to the purely cross-sectional results, it remains a selected factor in 10 of the 21 identified clusters.

Figure 4: Tree Cluster with Regimes

The tree is constructed to partition individual stock returns based on cross-sectional and time-series dimensions, using monthly data from 1980 to 2010. The initial two splits select aggregate predictors, namely liquidity and net stock issue, based on their standardized values within a 10-year rolling window, ranging from 0 to 1.



**Uncommon Factors** In addition to the common factors, there also exist some uncommon factors that are present in certain regimes. For example, Momentum (MOM) is the uncommon factor of the "Liquid" regime (Regime 1). It is selected with high probability only when market liquidity is low, but vanishes or crashes during liquid states. Momentum is often characterized as a behavioral anomaly that requires active arbitrage capital to exploit; during illiquid times, arbitrageurs face higher transaction costs, allowing the behavioral mispricing to persist. The Value (BM) premium is structurally confined to the "liquid and High Issuance" regime (Regime 3). This state is defined by high aggregate liquidity and high corporate financing activity. In this environment, the Value factor emerges as a prominent pricing driver. This is consistent with investment-based asset pricing theories: when firms are actively issuing equity (signaling high investment or distress financing) in a liquid market, the risk differential between value and growth widens.

### 5.3 Long-short Cluster Portfolio under Regime-Switching

Similarly to the cross-sectional investment, we also investigate the long-short cluster portfolio based on model prediction out-of-sample. The key distinction lies in the fact that different clusters may represent different time periods, both in-sample and out-of-sample. Thus, we form portfolios within each regime and aggregate these portfolios over time. It should be noted that at time t, we know the macro variable, so the next period can be classified into one specific regime, without introducing a look-ahead bias.

Table 13 shows model-predicted long-short portfolios under regime switching. These results are consistent with the cross-sectional evidence presented earlier. Focusing on the value-weighted case, the cluster portfolio with heterogeneous risk premia (the third column of Panel A) again dominates those who ignore clustering or assume homogeneous risk premia.

To determine whether the gains from our regime-dependent strategies are distinct from existing risk factors, we again perform spanning regressions, as shown in Table 14. The results confirm that the performance of the regime-switching strategy is not

Table 11: Regime-specific Factor Selection and Risk Premia Estimates

This table presents the Bayesian estimate of factor risk premia for each leaf cluster, computed from 2,000 MCMC samples, with the leaf cluster IDs matching those of the tree in Figure 4. The selection probability (in percent) is also presented in parentheses.

| Factor | N8 | N18 | N19 | N20 | N21 | N44 | N45 | N46 | N47 |
|---|---|---|---|---|---|---|---|---|---|
| | Regime 1: X_LIQ <0 | | | | | | | | |
| ME | 0.00(0) | 0.10(57) | -0.20(84) | 0.00(0) | -0.00(0) | 0.21(100) | -0.04(100) | -0.01(0) | -0.25(100) |
| ABR | 0.01(0) | 0.00(0) | 0.01(0) | 0.00(0) | 0.01(0) | 0.00(0) | 0.00(0) | 0.02(0) | 0.08(0) |
| SUE | 0.01(0) | 0.01(0) | 0.01(0) | 0.01(0) | 0.01(0) | 0.01(0) | 0.03(2) | 1.01(100) | 1.13(100) |
| BAS | -0.00(0) | -0.01(0) | -0.00(0) | -0.01(0) | -0.00(0) | -0.00(0) | -0.00(0) | -0.01(0) | -0.03(0) |
| MOM | 0.13(100) | 0.00(0) | -0.26(100) | 0.00(1) | 0.06(100) | 0.03(37) | -0.07(100) | -0.24(100) | -0.22(100) |
| BM | 0.00(0) | -0.00(0) | 0.00(0) | 0.00(0) | 0.00(0) | 0.00(0) | 0.21(100) | 0.01(0) | 0.06(0) |
| EP | 0.20(100) | 0.00(0) | 0.00(0) | -0.00(0) | 0.21(100) | 0.00(0) | 0.01(0) | 0.01(0) | 0.00(0) |
| CFP | 0.00(0) | 0.00(0) | 0.67(100) | 0.00(0) | 0.00(0) | 0.00(0) | 0.01(0) | 0.02(0) | 0.21(88) |
| SP | 0.00(0) | 0.00(0) | 0.00(0) | 0.00(0) | -0.00(0) | -0.00(0) | -0.02(100) | 0.01(0) | 0.36(100) |
| AGR | -0.06(100) | -0.00(0) | -0.09(100) | -0.00(0) | -0.00(0) | -0.00(0) | -0.01(0) | -0.01(0) | -0.09(0) |
| NI | -0.00(0) | -0.00(0) | -0.00(0) | -0.00(0) | -0.00(0) | -0.00(0) | -0.00(0) | -0.00(0) | -0.06(0) |
| ACC | -0.00(0) | -0.00(0) | -0.00(0) | -0.00(0) | -0.01(0) | -0.00(0) | -0.01(0) | -0.01(0) | -0.07(0) |
| OP | 0.00(0) | 0.28(100) | -0.00(0) | 0.00(0) | -0.00(0) | 0.27(85) | 0.00(0) | 0.01(0) | 0.02(0) |
| ROE | 0.00(0) | 0.00(0) | 0.61(100) | 0.38(100) | 0.41(100) | 0.00(0) | 0.01(0) | 0.00(0) | 0.01(0) |
| SEAS | -0.00(0) | 0.00(0) | 0.00(0) | 0.00(0) | -0.00(0) | 0.00(0) | 0.01(0) | 0.01(0) | 0.04(0) |
| ADM | -0.00(0) | 0.00(0) | 0.00(0) | 0.00(0) | 0.00(0) | 0.01(30) | 0.00(0) | 0.00(0) | 0.00(0) |
| RDM | 0.00(0) | 0.00(0) | -0.00(0) | 0.00(0) | 0.00(0) | 0.00(0) | 0.00(0) | 0.01(0) | 0.11(0) |
| SVAR | -0.00(0) | -0.69(100) | -0.55(100) | -0.00(0) | 0.02(100) | -0.00(0) | -0.00(0) | 0.18(100) | -0.54(100) |
| BETA | -0.26(100) | 0.25(100) | 0.12(100) | 0.18(99) | -0.12(100) | 0.00(0) | 0.13(100) | 0.08(100) | 0.12(100) |
| STR | -1.07(100) | -0.01(0) | -0.59(100) | -0.01(0) | -0.83(100) | 0.00(0) | -0.51(100) | -1.08(100) | -1.12(100) |

| | N24 | N25 | N52 | N53 | N54 | N55 | N28 | N29 | N60 | N61 | N62 | N63 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Regime 2: X_LIQ >0 & X_NI <0 | | | | | | Regime 3: X_LIQ >0 & X_NI >0 | | | | | |
| ME | 0.00(0) | -0.00(0) | 0.00(0) | -0.18(100) | -0.23(100) | -0.29(100) | 0.00(0) | 0.03(100) | 0.00(0) | -0.43(100) | 0.36(100) | -0.43(100) |
| ABR | 0.00(0) | 0.01(0) | 0.00(0) | 0.01(0) | 0.01(0) | 0.05(0) | 0.00(0) | 0.01(0) | 0.00(0) | 0.01(0) | 0.01(0) | 0.08(0) |
| SUE | 0.01(0) | 0.03(0) | 0.01(0) | 0.03(1) | 1.14(100) | 1.06(100) | 0.00(0) | 0.02(0) | 0.41(100) | 0.01(0) | 0.01(0) | 0.94(100) |
| BAS | -0.00(0) | -0.01(0) | -0.00(0) | -0.00(0) | -0.00(0) | -0.03(0) | -0.00(0) | -0.00(0) | -0.00(0) | -0.00(0) | -0.01(0) | -0.02(0) |
| MOM | 0.01(0) | 0.01(0) | 0.00(0) | 0.01(0) | 0.01(0) | 0.21(100) | 0.01(0) | 0.02(0) | 0.00(0) | 0.01(0) | 0.01(0) | 0.32(100) |
| BM | 0.00(0) | 0.00(0) | 0.00(0) | 0.00(0) | 0.02(0) | 0.03(0) | 0.66(100) | 0.05(100) | 0.00(0) | 0.00(0) | 0.47(100) | 0.53(100) |
| EP | 0.00(0) | 0.01(0) | 0.74(100) | 0.01(0) | 0.02(0) | 0.04(0) | 0.00(0) | 0.71(100) | 0.33(100) | 0.52(100) | 0.00(0) | 0.01(0) |
| CFP | 0.02(2) | 0.01(0) | 0.00(0) | 0.01(0) | 0.03(0) | 0.05(0) | 0.00(0) | 0.00(0) | 0.00(0) | 0.00(0) | 0.01(0) | 0.01(0) |
| SP | 0.00(0) | 0.00(0) | 0.00(1) | 0.00(0) | 0.01(0) | 0.28(100) | 0.00(0) | 0.00(0) | 0.00(0) | 0.00(0) | -0.01(0) | 0.30(100) |
| AGR | -0.00(0) | -0.00(0) | -0.00(0) | 0.00(0) | -0.01(0) | -0.01(0) | -0.00(0) | -0.00(0) | -0.00(0) | -0.00(0) | -0.03(0) | -0.05(0) |
| NI | -0.01(0) | -0.01(0) | -0.00(0) | -0.01(0) | -0.01(0) | -0.03(0) | 0.00(0) | -0.01(0) | -0.00(0) | -0.01(0) | -0.01(0) | -0.03(0) |
| ACC | -0.00(0) | -0.01(0) | -0.00(0) | -0.01(0) | -0.01(0) | -0.05(0) | 0.00(0) | -0.01(0) | -0.00(0) | -0.00(0) | -0.03(0) | -0.04(0) |
| OP | 0.01(0) | 0.01(0) | 0.00(0) | 0.01(0) | 0.02(0) | 0.04(0) | 0.00(0) | 0.01(0) | -0.00(0) | 0.00(0) | 0.00(0) | 0.36(100) |
| ROE | 0.01(0) | 0.01(0) | 0.01(0) | 0.01(0) | 0.01(0) | 0.03(0) | 0.46(100) | 0.01(0) | -0.00(0) | 0.00(0) | 0.00(0) | 0.04(0) |
| SEAS | -0.00(0) | 0.01(0) | 0.00(0) | 0.01(0) | 0.01(0) | 0.01(0) | 0.00(0) | 0.00(0) | 0.00(0) | 0.00(0) | 0.01(0) | 0.03(0) |
| ADM | 0.00(0) | -0.00(0) | -0.00(0) | -0.00(0) | -0.00(0) | 0.01(0) | -0.00(0) | -0.00(0) | 0.00(0) | -0.00(0) | -0.00(0) | 0.00(0) |
| RDM | 0.00(0) | 0.01(0) | 0.00(0) | 0.00(0) | 0.00(0) | 0.07(12) | 0.01(0) | 0.01(0) | 0.01(0) | 0.01(0) | 0.04(0) | 0.10(0) |
| SVAR | -0.01(0) | -0.51(100) | -0.00(0) | -0.44(100) | -0.49(100) | -0.35(100) | -0.00(0) | -0.39(100) | -0.00(0) | -0.25(100) | -0.40(100) | -0.05(100) |
| BETA | -0.00(0) | -0.14(100) | 0.00(0) | 0.00(0) | -0.00(0) | 0.04(85) | 0.00(0) | 0.00(0) | -0.00(0) | 0.00(0) | 0.01(0) | 0.13(100) |
| STR | 0.00(0) | -0.01(0) | -0.00(0) | -0.01(0) | -0.02(0) | -0.46(100) | -0.01(0) | -0.01(0) | -0.01(0) | -0.01(0) | -0.85(99) | -0.66(100) |

spanned by standard factors. Across all specifications, the portfolio delivers a statistically significant alpha, ranging from 0.49% to 0.61% per month ($t$-stats between 2.03 and 3.12). Notably, in model 1, the strategy yields a monthly alpha of 0.60% ($t$-stat = 3.10) after controlling for the FF5 factors and Momentum. The factor loadings reveal that while the strategy has some exposure to value (HML) and profitability (RMW), these exposures do not fully explain its returns. Overall, the significant unexplained alpha indicates that incorporating macroeconomic regimes into the BFCM framework captures dynamic, state-dependent pricing information that common factor models fail to identify.

It appears that the investment performance under the regime-switching framework is lower than that of the pure cross-section. This discrepancy can be attributed

Table 12: Regime Switching: Factor Selection Evaluation by Leaf Clusters

This table presents the uncommon factor selection outcomes for the cross-sectional clustered model during regime changes in Figure 4. The rows labeled $> 0.95$ correspond to selection probability thresholds for the factors. The order of the listed factors is based on their selection probability values. Three panels are included to show the leaf cluster positions for the three macroeconomic-targeted regimes.

| | N8 | N18 | N19 | N20 | N21 | N44 | N45 | N46 | N47 |
|---|---|---|---|---|---|---|---|---|---|
| | Regime 1: X_LIQ <0 | | | | | | | | |
| >95% | EP AGR STR MOM BETA | OP SVAR BETA | MOM CFP AGR ROE SVAR BETA STR | ROE BETA | MOM EP ROE SVAR BETA STR | ME | ME MOM BM SP BETA STR | SUE MOM SVAR BETA STR | ME SUE MOM SP SVAR BETA STR |

| | N24 | N25 | N52 | N53 | N54 | N55 | | N28 | N29 | N60 | N61 | N62 | N63 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Regime 2: X_LIQ >0 & X_NI <0 | | | | | | Regime 3: X_LIQ >0 & X_NI >0 | | | | | |
| >95% | | SVAR BETA | EP | ME SVAR | ME SUE SVAR | ME SUE MOM SP SVAR STR | | BM ROE | ME BM EP SVAR | SUE EP | ME EP SVAR | ME BM SVAR STR | ME SUE MOM BM SP OP SVAR BETA STR |

to the inherent challenge of modeling dynamic regimes versus stable cross-sectional patterns. The regime-switching framework introduces time-series partitions which inevitably reduce the effective sample size for estimating factor premia within each macroeconomic state. As macroeconomic conditions fluctuate, capturing these transient time-series dynamics proves more prone to estimation error than identifying cross-sectional heterogeneity.

Table 13: Model Predicted Long Short Portfolios under Regime Switching

This table presents the long-short portfolios based on model prediction. The cluster portfolio forms long-short (20%) portfolios in each cluster and uses the in-sample (1980-2010) estimate to calculate the tangency portfolio weight, then applies it to out-of-sample (2011-2024) cluster portfolios. The overall portfolio forms long-short portfolios based on the full cross-section. We consider both equal-weighted (EW) and value-weighted (VW). Panel A uses cluster-wise coefficient estimates, while Panel B uses overall pool estimates.

| | Cluster EW | Overall EW | Cluster VW | Overall VW |
|---|---|---|---|---|
| | Panel A: Bayes Coef | | | |
| Mean Return | 3.13 | 2.26 | 0.66 | 0.37 |
| $t$-stat | (11.81) | (11.53) | (2.94) | (1.29) |
| Median Return | 2.92 | 2.36 | 0.47 | 0.39 |
| SD | 3.95 | 3.03 | 2.87 | 3.03 |
| Skewness | 0.75 | 0.21 | 0.49 | 0.57 |
| Kurtosis | 2.41 | 1.13 | 1.35 | 4.33 |
| AC1 | -0.06 | -0.14 | 0.04 | 0.14 |
| Sharpe Ratio | 2.74 | 2.58 | 0.80 | 0.42 |
| Max Drawdown | 0.09 | 0.12 | 0.16 | 0.35 |
| | Panel B: Bayes Pool Coef | | | |
| Mean Return | 2.39 | 2.17 | 0.55 | 0.41 |
| $t$-stat | (9.28) | (9.67) | (1.98) | (1.61) |
| Median Return | 2.22 | 2.31 | 0.42 | 0.56 |
| SD | 3.35 | 2.89 | 3.36 | 2.80 |
| Skewness | 0.47 | 0.12 | 0.96 | 0.34 |
| Kurtosis | 2.63 | 0.46 | 3.53 | 2.26 |
| AC1 | -0.07 | -0.03 | -0.03 | 0.09 |
| Sharpe Ratio | 2.47 | 2.61 | 0.57 | 0.51 |
| Max Drawdown | 0.13 | 0.11 | 0.21 | 0.30 |

Table 14: Spanning Regression of Cluster Long-short Portfolio under Regime Switching

This table presents the spanning regression of the Bayes Coef Cluster VW portfolio in Table 13 on several factor models. The factors include FF5 factor, momentum factor (UMD), liquidity factor (LIQ) from Pástor and Stambaugh (2003), q-factor from Hou et al. (2015), quality minus junk (QMJ) from Asness et al. (2019), betting-against-beta (BAB) from Frazzini and Pedersen (2014).

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Alpha | 0.601*** (3.099) | 0.611*** (3.115) | 0.581** (2.564) | 0.564** (2.579) | 0.490** (2.031) | 0.560*** (2.798) |
| MktRf | 0.003 (0.059) | -0.002 (-0.041) | 0.060 (1.137) | 0.069 (1.239) | 0.099* (1.665) | 0.007 (0.122) |
| SMB | 0.075 (0.876) | 0.068 (0.782) | | | | 0.383 (1.213) |
| HML | 0.461*** (5.673) | 0.452*** (5.330) | | | | 0.426*** (4.550) |
| RMW | 0.405*** (3.847) | 0.399*** (3.753) | | | | 0.415*** (2.704) |
| CMA | -0.139 (-1.147) | -0.130 (-1.047) | | | | -0.942*** (-3.746) |
| UMD | | -0.024 (-0.390) | | | | -0.049 (-0.688) |
| LIQ | | | -0.154** (-2.337) | | | 0.012 (0.199) |
| ME | | | | 0.100 (1.042) | | -0.297 (-0.957) |
| IA | | | | 0.378*** (3.661) | | 0.848*** (3.808) |
| ROE | | | | 0.253** (2.297) | | 0.203 (1.322) |
| EG | | | | -0.159 (-1.164) | | 0.045 (0.309) |
| QMJ | | | | | 0.208** (2.115) | -0.181 (-1.248) |
| BAB | | | | | -0.017 (-0.179) | 0.039 (0.444) |

## 6 Conclusion

We revisit the challenge of dimensionality in empirical asset pricing through a lens of market segmentation and grouped heterogeneity. While current literature primarily focuses on taming the "factor zoo" through global sparsity – assuming a single set of factors applies to all assets – we demonstrate that this approach obscures the fundamental heterogeneity of financial markets. By forcing a "one-size-fits-all" model onto the cross-section, researchers risk discarding signals that are locally significant but globally weak.

Our empirical investigation yields three substantive insights for the future of asset pricing research. First, we provide robust evidence of heterogeneous risk premia across U.S. equities. We show that the pricing of risk is not universal but is instead structured around asset clusters defined by idiosyncratic volatility, market capitalization, and macroeconomic regimes. Second, we introduce a new taxonomy of risk factors, distinguishing between pervasive "common factors" and "uncommon factors" that command significant premia only within specific localized subsets. This distinction helps resolve long-standing debates regarding the apparent instability of established factors like Value and Momentum.

Finally, we demonstrate that this heterogeneity represents a distinct investment opportunity. Out-of-sample portfolios constructed using cluster-specific model generate Sharpe ratios significantly higher than those produced by traditional pooled models. These results confirm that the "factor zoo" is dynamic, state-dependent, and inherently localized, suggesting that the path forward for empirical asset pricing lies in models that move beyond universal assumptions toward a more granular understanding of market structure.

## References

Ahn, D.-H., J. Conrad, and R. F. Dittmar (2009). Basis assets. *Review of Financial Studies 22*(12), 5133–5174.

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

Avramov, D., S. Cheng, L. Metzker, and S. Voigt (2023). Integrating factor models. *Journal of Finance 78*(3), 1593–1646.

Avramov, D. and T. Chordia (2006). Asset pricing models and financial market anomalies. *Review of Financial Studies 19*(3), 1001–1040.

Barillas, F. and J. Shanken (2018). Comparing asset pricing models. *Journal of Finance 73*(2), 715–754.

Bekaert, G., R. J. Hodrick, and X. Zhang (2009). International stock return comovements. *Journal of Finance 64*(6), 2591–2626.

Bianchi, D., M. Büchner, and A. Tamoni (2021). Bond risk premiums with machine learning. *Review of Financial Studies 34*(2), 1046–1089.

Bonhomme, S. and E. Manresa (2015). Grouped patterns of heterogeneity in panel data. *Econometrica 83*(3), 1147–1184.

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

Bryzgalova, S., M. Pelger, and J. Zhu (2025). Forest through the trees: Building cross-sections of stock returns. *The Journal of Finance 80*(5), 2447–2506.

Chaieb, I., H. Langlois, and O. Scaillet (2021). Factors and risk premia in individual international stock returns. *Journal of Financial Economics 141*(2), 669–692.

Chib, S. and X. Zeng (2020). Which factors are risk factors in asset pricing? A model scan framework. *Journal of Business & Economic Statistics 38*(4), 771–783.

Chib, S., L. Zhao, and G. Zhou (2023). Winners from winners: A tale of risk factors. *Management Science, Forthcoming*.

Chipman, H. A., E. I. George, and R. E. McCulloch (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics 4*, 266–298.

Cong, L., G. Feng, J. He, and X. He (2025). Growing the Efficient Frontier on Panel Trees. *Journal of Financial Economics 167*, 104024.

Cong, L. W., G. Feng, J. He, and J. Li (2024). Sparse modeling under grouped heterogeneity with applications to asset pricing. Technical report, City University of Hong Kong.

Cui, L., G. Feng, J. Ma, and Y. Su (2025). Breaks and trends in factor premia. Technical report, City University of Hong Kong.

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

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

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

Fama, E. F. and K. R. French (2020). Comparing cross-section and time-series factor models. *The Review of Financial Studies 33*(5), 1891–1926.

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

Foerster, S. R. and G. A. Karolyi (1999). The effects of market segmentation and investor recognition on asset prices: Evidence from foreign stocks listing in the United States. *Journal of Finance 54*(3), 981–1013.

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

George, E. I. and R. E. McCulloch (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association 88*(423), 881–889.

Giannone, D., M. Lenza, and G. E. Primiceri (2021). Economic predictions with big data: The illusion of sparsity. *Econometrica 89*(5), 2409–2437.

Griffin, J. M. (2002). Are the Fama and French factors global or country specific? *Review of Financial Studies 15*(3), 783–803.

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

Harvey, C. R. (2017). Presidential Address: The Scientific Outlook in Financial Economics. *Journal of Finance 72*, 1399–1440.

He, J. and P. R. Hahn (2021). Stochastic tree ensembles for regularized nonlinear regression. *Journal of the American Statistical Association*, 1–20.

He, J., S. Yalov, and P. R. Hahn (2019). XBART: Accelerated Bayesian additive regression trees. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1130–1138.

Hou, K., G. A. Karolyi, and B.-C. Kho (2011). What factors drive global stock returns? *Review of Financial Studies 24*(8), 2527–2574.

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

Hwang, S. and A. Rubesam (2020). Bayesian Selection of Asset Pricing Factors Using Individual Stocks. *Journal of Financial Econometrics*.

Jarrow, R. A., R. Murataj, M. T. Wells, and L. Zhu (2020). The low-volatility anomaly and the adaptive multi-factor model. *arXiv preprint arXiv:2003.08302*.

Kan, R., C. Robotti, and J. Shanken (2013). Pricing model performance and the two-pass cross-sectional regression methodology. *Journal of Finance 68*(6), 2617–2649.

Karolyi, G. A. and R. M. Stulz (2003). Are financial assets priced locally or globally? *Handbook of the Economics of Finance 1*, 975–1020.

Kozak, S., S. Nagel, and S. Santosh (2020). Shrinking the cross-section. *Journal of Financial Economics 135*(2), 271–292.

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

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

Patton, A. J. and B. M. Weller (2022). Risk price variation: The missing half of empirical asset pricing. *Review of Financial Studies 35*(11), 5127–5184.

Rossi, A. G. and A. Timmermann (2015). Modeling covariance risk in Merton's ICAPM. *Review of Financial Studies 28*(5), 1428–1461.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley.

Smith, S. C. and A. Timmermann (2021). Break risk. *Review of Financial Studies 34*(4), 2045–2100.

Smith, S. C. and A. Timmermann (2022). Have risk premia vanished? *Journal of Financial Economics 145*(2), 553–576.

Van Binsbergen, J. H., X. Han, and A. Lopez-Lira (2023). Man versus machine learning: The term structure of earnings expectations and conditional biases. *Review of Financial Studies 36*(6), 2361–2396.

Welch, I. and A. Goyal (2008). A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies 21*(4), 1455–1508.

# Online Appendices

## I Proof of Proposition 1

*Proof.* To simplify notations, we re-express the model in (2) as

$$\mathbf{R}_t = \mathbf{1}\mathrm{R}_{z,t} + \mathbf{Z}_{t-1}\mathbf{f}_t + \epsilon_{i,t} = \mathcal{W}_t\boldsymbol{\beta}_t + \boldsymbol{\epsilon}$$

where $\mathcal{W}_t = [\mathbf{1}, \mathbf{Z}_{t-1}]$.

First, we keep the conditions with respect to the assignment of $\boldsymbol{\gamma}_j$ to spike or slab and integrate out all other parameters. Given $\boldsymbol{\gamma}_j$, the prior on $\boldsymbol{\beta}$ is $N(0, \sigma_j^2\boldsymbol{\Lambda}_{0|\boldsymbol{\gamma}_j}^{-1})$, where $\boldsymbol{\Lambda}_{0|\boldsymbol{\gamma}_j}$ is a diagonal matrix with $s$-th element being $\xi_1^{-2}\sigma_j^{-2}$ if $\gamma_{j,s} = 1$, or $\xi_0^{-2}\sigma_j^{-2}$ if $\gamma_{j,s} = 0$. Following the standard result of Bayesian linear regression with conjugate Normal-Inverse-Gamma prior, integrating out all regression coefficients and residual variance yields,

$$p(\mathbf{R}_t \mid \boldsymbol{\gamma}_j, \mathbf{Z}_{t-1}) = \frac{1}{(2\pi)^{N/2}}\sqrt{\frac{|\boldsymbol{\Lambda}_{0|\boldsymbol{\gamma}_j}|}{|\boldsymbol{\Lambda}_N|}}\frac{v_0^{S_0}}{v_N^{S_N}}\frac{\Gamma(S_N)}{\Gamma(S_0)},$$

where $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}dx$ is the gamma function.

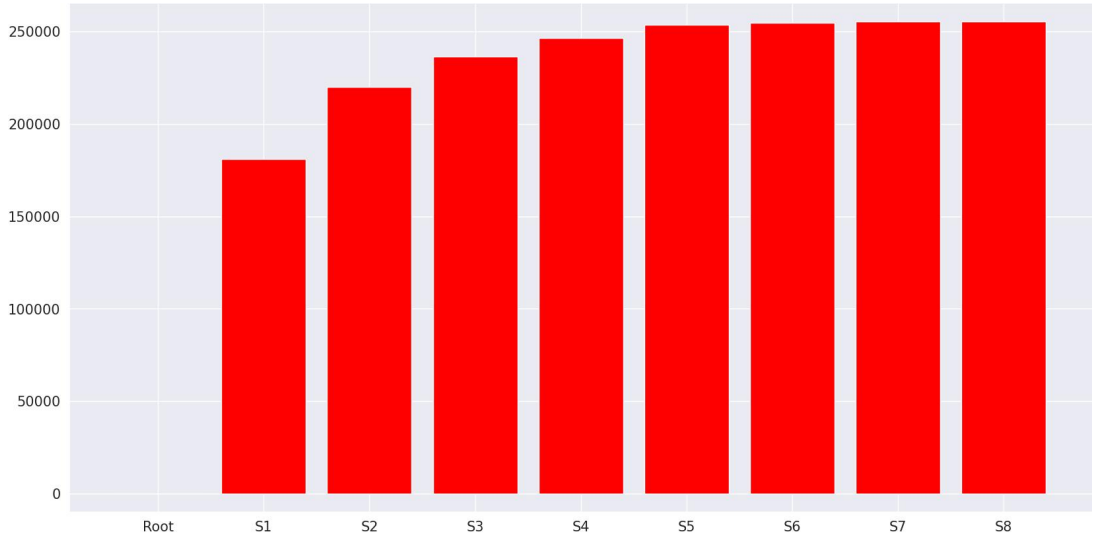
Second, we integrate out the indicator of assignment $\boldsymbol{\gamma}_j$, and sum across all time period $T$. Let $\{I_m\}_{m=1}^{2^K}$ denote a collection of all length $K$ vectors with each element taking values 0 or 1, indicating all possible values that $\boldsymbol{\gamma}_{j,\mathbf{f}}$ can take. Note the prior of each model is $\pi(\boldsymbol{\gamma}_{j,\mathbf{f}} = I_m) = \prod_{k=1}^K w_k^{\gamma_{j,\mathbf{f},k}}(1-w_k)^{(1-\gamma_{j,\mathbf{f},k})}$ is defined in (4). The Bayes rule yields,

$$p(\mathbf{R} \mid \mathbf{Z}) = \prod_{t=1}^T \sum_{I_m} \frac{1}{2^K}\pi(\boldsymbol{\gamma}_{j,\mathbf{f}} = I_m)p(\mathbf{R}_t \mid \boldsymbol{\gamma}_j, \mathbf{Z}_{t-1}).$$

□

# II Figure: Marginal (Log) Likelihood Improvements

Figure A.1: This bar plot shows the increase in log marginal likelihood (without normalizing constants) compared to the root model with no splits. To avoid overfitting, the tree applies a penalty (12) during growth and stops after 8 splits, resulting in 9 terminal leaves.



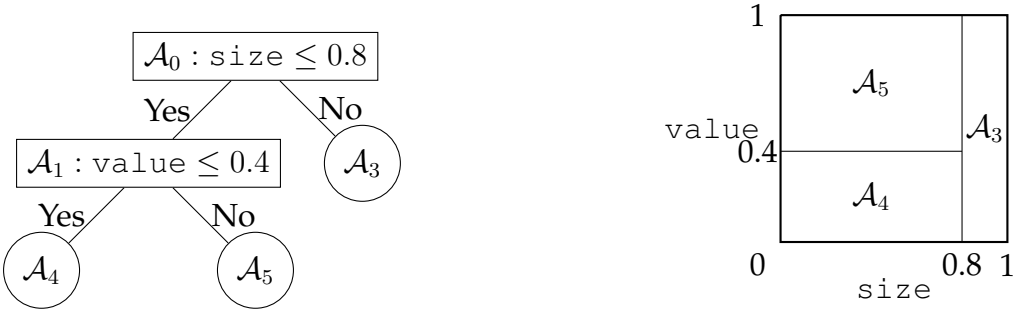

## II.1 Tree Clustering Procedure

With the knowledge of how we select factors (and models) in each cluster, we discuss how the clusters are jointly determined. We first introduce a global split criterion for growing the tree based on the marginal likelihood of the model in equation (2), and then describe the Gibbs sampler for inferring parameters in the model. A Tree is essentially a sequence of split rules (cutpoints) that partitions the space of characteristics into many hyper-rectangular regions (leaf nodes). By splitting based on firm characteristics, the BFCM tree can learn cross-sectional heterogeneity as assets with different characteristic values are partitioned into distinct sides. For simplicity, the following section expresses a split rule $\{z \leq c\}$ as $\tilde{c}$. The split rule $\{z \leq c\}$ at each intermediate node of the tree checks whether a specific variable ($z$) is smaller than a threshold ($c$). Figure A.2 illustrates.

Suppose a new asset return with characteristics (e.g., size and value) comes; we start by passing it from the (top) root node $\mathcal{A}_0$, check all the split rules, and then assign it to one and only one of the three (bottom) *leaf nodes*, or *clusters*. All leaf nodes con-

Figure A.2: **Decision Tree Partitioning Two Dimensional Characteristics Space.**

Left: A decision tree with two splits and three leaves. Right: Corresponding partition plot for the predictor space spanned by `size` and `value`. This tree partitions the entire space into three rectangular $\mathcal{A}_3$, $\mathcal{A}_4$, $\mathcal{A}_5$.




stitute a partition of the entire space, as shown in the right panel. We refer to nodes other than the root and final leaves as intermediate nodes (with split rules), such as $\mathcal{A}_1$ in Figure A.2. $J$ is then the number of final leaves.

A BFCM tree grows iteratively. In the baseline specification, before the first split, the entire cross section of the asset returns is at the root of the tree, $\mathcal{A}_0$. BFCM starts from the root node, searches for the optimal split rule that maximizes the objective (the split criterion), among all possible candidates defined by various characteristics and thresholds, and creates two leaf nodes (bottom nodes). Then we evaluate the leaf nodes to find the second optimal split rule. Besides the split rule candidates, we add one more option of "stop splitting" at each iteration, with a penalty of the tree size to help the tree stop once there is no desirable split rule. It prevents the tree from growing too large (thus making it less interpretable) and alleviates the problem of overfitting. The iteration terminates once all leaf nodes reach the "stop splitting" condition or other stopping criteria, such as reaching the pre-specified maximum depth of the tree or the minimum number of data observations in a leaf node. We next elaborate on the procedures and include the pseudo codes 1:

### II.1.1 First Split

We explore all split rule candidates, including every pair of firm characteristics and split thresholds, to partition the cross-section (panel). All characteristics are normalized cross-sectionally to the range $[-1, 1]$ based on percentiles. Figure A.3 presents

**Algorithm** Bayesian Fama-Macbeth Clustering Model

1: **procedure** BFCM
2: **Input:** data and parameters. **Output:** a tree structure with a bunch of split rules that define clusters by panel characteristics.
3: **for** $j$ from 1 to num_iter **do** ▷ Loop over number of iterations
4: **if** current depth $\geq d_{\max}$ **then**
5: **return**.
6: **else**
7: Search the tree, find all leaf nodes $\mathcal{A}$
8: **for** each leaf node $s$ in $\mathcal{A}$ **do** ▷ Loop over all current leaf nodes
9: **if** $s$ is not labeled as "cannot split" **then**
10: **for** each split candidate $\tilde{c}_m$ in $\mathcal{C}_s$ **do**
11: Partition data temporally in $N$ according to $\tilde{c}_m$.
12: **if** Left or right child node cannot satisfy minimal leaf size **then**
13: $L(\tilde{c}_m) = -\infty$.
14: **else**
15: Calculate the split criteria $L(\tilde{c}_m)$ in (13).
16: **end if**
17: **end for**
18: **end if**
19: **end for**
20: Find the best leaf node and split rule that maximizes split criteria

$$\tilde{c}_j = \underset{N\in\mathcal{N},\tilde{c}_m\in\cup_{s=1}^{|\mathcal{N}|}\mathcal{C}_s}{\arg\max} \{L(\tilde{c}_j)\}$$

21: Split the node selected at the $j$-th split rule of the tree $\tilde{c}_j$. If a null cutpoint is selected, label the corresponding node "cannot split" and return to line 7.
22: **end if**
23: **end for**
24: **for** $j$ from 1 to $J$ **do** ▷ Loop over all clusters returned by the tree
25: Draw posterior samples of parameters by the Gibbs sampler in Section 3.3.
26: **end for**
27: **return**
28: **end procedure**

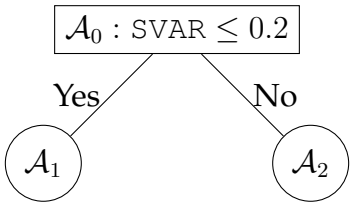
Note that the number of clusters $J$ is not determined by users, but is learnt from data after the `for` loop ending in line 23.

one split rule candidate, which partitions the sample into the left and right child leaf nodes, $\mathcal{A}_1$ and $\mathcal{A}_2$, respectively, based on whether an individual asset's `SVAR` falls in the bottom 60 (from -1 to 0.2) percentiles.

The partitioning generally aims to group similar observations into the same leaf to fit a locally sparse model well. To evaluate split rule candidates, the "fitness" of the resulting factor model at each leaf is a natural split criterion. However, the commonly used goodness of fitness, likelihood function, of the model in (2) involves unknown parameters that cannot be accurately estimated given the noisy data, which may favor a bad split rule. Instead, we use the closed-form expression of the *marginal likelihood*,

Figure A.3: **Illustration of one candidate for the first split**

To calculate the split criterion and search for the best characteristic to split at the optimal cutpoint, let us consider one split candidate, $\texttt{SVAR} \leq 0.2$.




where all fitted parameters are integrated out, to address any concerns about parameter uncertainty during tree growth.

**Marginal likelihood.** At each month $t$, stack all data in the *same cluster* in matrix form, $\mathbf{R}_t = [r_{1,t}, ..., r_{n,t}]^\mathsf{T}$, $\mathbf{Z}_{t-1} = [\mathbf{z}_{1,t-1}, ..., \mathbf{z}_{n,t-1}]^\mathsf{T}$, then the marginal likelihood of the model at node $\mathcal{A}_0$ for month $t$ is given by:

$$\begin{aligned} p(\mathcal{A}_0) := \prod_{t=1}^{T} p(\mathbf{R}_t \mid \mathbf{Z}_{t-1}) = \prod_{t=1}^{T} \int p(\mathbf{R}_t \mid \mathbf{Z}_{t-1}, \boldsymbol{\gamma}_j, r_{j,z,t}, \mathbf{f}_j, \sigma_j^2) \\ \times \pi(r_{j,z,t} \mid \sigma_{j,t}^2) \pi(\mathbf{f}_{j,t} \mid \sigma_j^2, \boldsymbol{\gamma}_j) \pi(\sigma_{j,t}^2 \mid \boldsymbol{\gamma}_j) \pi(\boldsymbol{\gamma}_j) dr_{j,z,t} d\mathbf{f}_{j,t} d\sigma_{j,t}^2 d\boldsymbol{\gamma}_j. \end{aligned} \tag{10}$$

Intuitively, the marginal likelihood takes the expectation of the unknown parameters in the likelihood function with respect to the prior distributions. A function of the data and prior parameters only, it accounts for parameter estimation and model selection uncertainties, separating the tree growth from factor model estimations.

**Split and stopping criteria.** We collect all split rule candidates in $\mathcal{C} = \{\tilde{c}_j\}$, including candidates that split each variable at all possible cutpoint values. Recall that each split rule candidate $\tilde{c}_j$ partitions the current node into two child leaf nodes (Figure A.3) before we fit the factor model at each potential leaf cluster separately. Because $\epsilon_{i,t}$ are independent, the two child nodes are independent, and the joint marginal likelihood of the entire data is the product of the marginal likelihoods of the two child nodes,

$$l(c_j) = p(\mathbf{R}^L \mid \mathbf{Z}^L, \mathbf{F}^L) \times p(\mathbf{R}^R \mid \mathbf{Z}^R, \mathbf{F}^R) = p(\mathcal{A}_1) \times p(\mathcal{A}_2). \tag{11}$$

The superscript $L$ and $R$ represent data observations partitioned to the left and right child nodes by the split rule candidate $\tilde{c}_j$. Each split rule candidate partitions the data to $\mathcal{A}_1$ and $\mathcal{A}_2$ differently, and the value of (11) varies across split rule candidates, serving as a natural measurement of model fitness and quality of the candidate, with larger values better.

Furthermore, besides all the split rule candidates, we consider the option to stop splitting at the current node definitively, the *null cutpoint* option, the marginal likelihood of which is defined as:

$$l(\emptyset) = |\mathcal{C}| \left( \frac{(1+d)^{\tilde{a}_2}}{\tilde{a}_1} - 1 \right) p(\mathbf{R} \mid \mathbf{Z}, \mathbf{F}) = |\mathcal{C}| \left( \frac{(1+d)^{\tilde{a}_2}}{\tilde{a}_1} - 1 \right) p(\mathcal{A}_0), \tag{12}$$

where $|\mathcal{C}|$ is the number of all split rule candidates, $d$ is the depth of the current node (root node has depth 1), and $\tilde{a}_1$ and $\tilde{a}_2$ are two hyperparameters. The marginal likelihood in (12) is evaluated on *all* data in the root node, since all the data belong to the same current node if we stop splitting further. We follow the XBART framework (He et al., 2019; He and Hahn, 2021) to set the early stop option to match the regularization tree prior used in BART (Chipman, George, and McCulloch, 2010) that chooses $\tilde{a}_1 = 0.95$ and $\tilde{a}_2 = 2$. The additional weight on the null cutpoint criterion $|\mathcal{C}| \left( \frac{(1+d)^{\tilde{a}_2}}{\tilde{a}_1} - 1 \right)$ is chosen as regularization of the tree. As the tree grows, $d$ increases, and the weight (penalty) increases exponentially. If the benefit of splitting does not outweigh the growth penalty on tree size, the tree stops growing to avoid overfitting.

Once all marginal likelihoods have been calculated for all split rule candidates and the null cutpoint, assuming the equal prior probability of each split rule candidate, the split criterion can be derived following the Bayes rule:

$$L(c_j) = \frac{l(c_j)}{\sum_{j'} l(c_{j'}) + l(\emptyset)}, \quad L(\emptyset) = \frac{l(\emptyset)}{\sum_{j'} l(c_{j'}) + l(\emptyset)}. \tag{13}$$

The split criterion in (13) is essentially the posterior probability of a split candidate with the prior probability of each candidate proportional to 1. We choose the one that maximizes (13) as the first split rule. This is a maximum a posteriori (MAP) estimator,

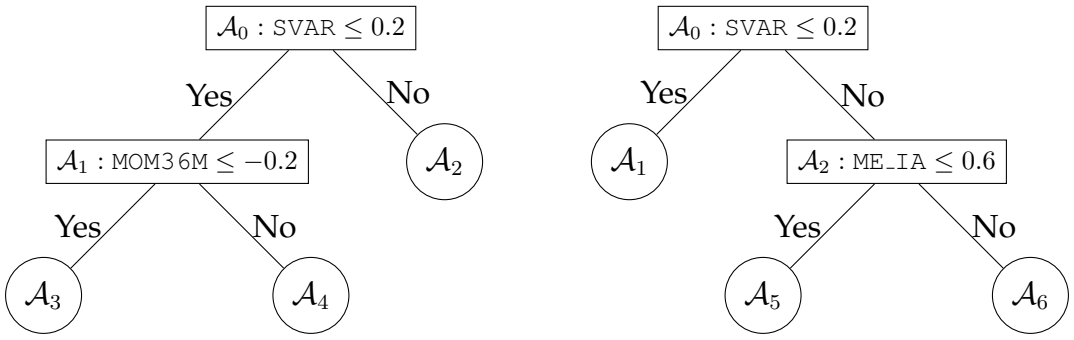
as it maximizes the posterior probability. Once a split rule is selected, we split the current node, yielding two child leaf nodes. However, if the null cutpoint is selected, the current node will not split further.

### II.1.2 Second Split

If the first split is not the null cutpoint, the second split can happen at either the left or right child node of the root node. Therefore, all split candidates of both nodes should be considered, and the split rule candidates expand to $\mathcal{C} = \{c_k^{\mathcal{A}_1}\} \cup \{c_k^{\mathcal{A}_2}\}$, where the superscript denotes which node to split. Figure A.4 illustrates two split rule candidates, for example, at either side of the root.

Figure A.4: **Illustration of the Second Split Candidates**

This figure illustrates two example candidates for the second split. Note that no matter which leaf node to split, the second ($j$-th) iteration of the tree algorithm has three ($j+1$) leaf nodes.




(a) If splitting node $\mathcal{A}_1$ at MOM3 6M.

(b) If splitting node $\mathcal{A}_2$ at ME_IA.

No matter which one is chosen, one leaf node splits into two new leaf nodes for the next iteration. The joint marginal likelihood corresponding to Eq. (11) is:

$$l(c_j^{\mathcal{A}_1}) = p(\mathcal{A}_3) \times p(\mathcal{A}_4) \times p(\mathcal{A}_2), \quad l(c_j^{\mathcal{A}_2}) = p(\mathcal{A}_1) \times p(\mathcal{A}_5) \times p(\mathcal{A}_6), \tag{14}$$

which is essentially the product of the marginal likelihood evaluated on *all* resulting leaves. The marginal likelihood of a null cutpoint is defined similarly:

$$l(\emptyset^{\mathcal{A}_1}) = |\mathcal{C}| \left( \frac{(1+d)^{\tilde{a}_2}}{\tilde{a}_1} - 1 \right) p(\mathcal{A}_1) \times p(\mathcal{A}_2)$$
$$l(\emptyset^{\mathcal{A}_2}) = |\mathcal{C}| \left( \frac{(1+d)^{\tilde{a}_2}}{\tilde{a}_1} - 1 \right) p(\mathcal{A}_2) \times p(\mathcal{A}_1),$$

where the null cutpoints $\emptyset^{\mathcal{A}_1}$ and $\emptyset^{\mathcal{A}_2}$ denote stop splitting the node $\mathcal{A}_1$ or $\mathcal{A}_2$, respectively. Note that $d = 2$ for the second split, since $\mathcal{A}_1$ or $\mathcal{A}_2$ has depth 2.

Similarly, for all candidates in $\mathcal{C} = \{c_j^{\mathcal{A}_1}\} \cup \{c_j^{\mathcal{A}_2}\}$ and two nulls, we have:

$$\begin{aligned} W &= \sum_{j'} l(c_{j'}^{\mathcal{A}_1}) + \sum_{j'} l(c_{j'}^{\mathcal{A}_2}) + l(\emptyset^{\mathcal{A}_1}) + l(\emptyset^{\mathcal{A}_2}) \\ L(c_j^{\mathcal{A}_1}) &= \frac{l(c_j^{\mathcal{A}_1})}{W}, L(\emptyset^{\mathcal{A}_1}) = \frac{l(\emptyset^{\mathcal{A}_1})}{W}, L(c_j^{\mathcal{A}_2}) = \frac{l(c_j^{\mathcal{A}_2})}{W}, L(\emptyset^{\mathcal{A}_2}) = \frac{l(\emptyset^{\mathcal{A}_2})}{W}. \end{aligned} \tag{15}$$

The second split is the one maximizing (15). We emphasize that the split criterion is defined *globally* — (14) is defined on all resulting leaf nodes and all data. Thus, the second split maximizes the global marginal likelihood to avoid overfitting. This global split criterion avoids the myopic local split criterion used in standard machine learning recursive algorithms. The node no longer splits when the null cutpoint is chosen for a specific node, say $\mathcal{A}_1$.

### II.1.3 Further Splits

The clustering algorithm proceeds iteratively by splitting leaf nodes based on the improvement in marginal likelihood. All existing leaf nodes are considered for each split, and the split rule candidates are evaluated. The global split criterion is defined on all resulting leaves, and the candidates that maximize the split criterion are chosen. The tree-growing process terminates once all leaf nodes have chosen the early stop option or some pre-specified stopping conditions are met. The tree creates $J$ leaf clusters if it terminates splitting after $J - 1$ iterations. It is worth repeating that by using marginal likelihood, we avoid intermediate model estimation errors. Yet, the choice of split rules, or tree growth, is still determined by Bayesian variable selection in each resulting cluster. Moreover, the joint marginal likelihood has a natural economic interpretation, whereas a global split criterion under a non-Bayesian framework has to be defined ad hoc.