Factor Identity

Jiantao Huang Christian Julliard Ran Shi*

March 2026

Abstract

The empirical dominance of dense factor models for pricing the cross-section of expected returns has come at a cost: the economic narrative of priced risk has disappeared. We introduce *factor identities*—groups of factors conveying the same pricing information despite modest correlations—and a Bayesian method to recover them, restoring legibility to the dense SDF. Only a handful emerge: a macroeconomic identity revealed through long-horizon growth in output, consumption, and industrial production; a market identity pinning down the level of risk premia; and a few characteristic-based identities that are fully subsumed by latent factors. The factor zoo is dense in animals but sparse in species.

Keywords: Asset pricing, factor models, factor identity, stochastic discount factor, macro-finance, Bayesian inference, variational approximation

JEL Classification: C11, C52, E44, G12

*Huang: Faculty of Business and Economics, University of Hong Kong, huangjt@hku.hk; Julliard: Department of Finance, FMG, and SRC, London School of Economics, and CEPR, c.julliard@lse.ac.uk; Shi: Leeds School of Business, University of Colorado Boulder, ran.shi@colorado.edu. For helpful comments, discussions, and suggestions, we thank Mike Chernov, Alex Dickerson, Valentin Haddad, Lars Lochstoer, Philippe Mueller, Avanidhar (Subra) Subrahmanyam, Ivo Welch, and Shiyang Huang.

"Things are the same of which one can be substituted for the other without loss of truth." G. W. Leibniz (1686)

A growing consensus holds that the stochastic discount factor is dense and complex: many factors contribute simultaneously to pricing the cross-section of expected returns, simple low-dimensional models are empirically dominated, and no small set of individually named factors suffices.[1] Yet this density comes at a cost. The benchmark models of the past, the canonical narratives that gave priced risk an economic face, are proven insufficient. And the path forward is equally obscured: when a new theory implies a new state variable, there is no principled way to ask whether it represents a genuinely new source of priced risk or merely a proxy for something already known. Against a dense model containing dozens of factors, a single new channel has no econometric chance of demonstrating distinctiveness, however well-motivated the theory. The economic narrative has been crowded out by empirical density.

We propose a way out. We introduce the concept of a *factor identity*: a group of factors that convey the same pricing information for the cross-section of expected returns,[2] even when their pairwise time-series correlations are modest. Two factors share an identity if the risk exposures they generate across assets are close substitutes: the data cannot distinguish their separate contributions to expected returns, and it is meaningless to designate either one alone as the driver of risk premia. Mapping the factor zoo into its underlying identities restores economic legibility to the SDF and provides the inferential apparatus against which new theories can be tested: a proposed state variable adds genuine economic content if and only if it introduces an identity not already spanned by the existing pricing groups.

Applying our method to a comprehensive set of macroeconomic variables, characteristic-sorted portfolios, and statistical latent factors, we uncover a striking result: the SDF is dense in animals but sparse in species. Only a handful of distinct economic identities drive the cross-section of expected returns: a macroeconomic identity, best captured by long-horizon growth in output, industrial production, and consumption; a market identity, which pins down the overall level of risk premia; and a latent statistical identity, reflecting residual variation not yet attributable to observable fundamentals. Characteristic-based factors do carry some genuine pricing information, best captured by a size identity and a momentum/mispricing

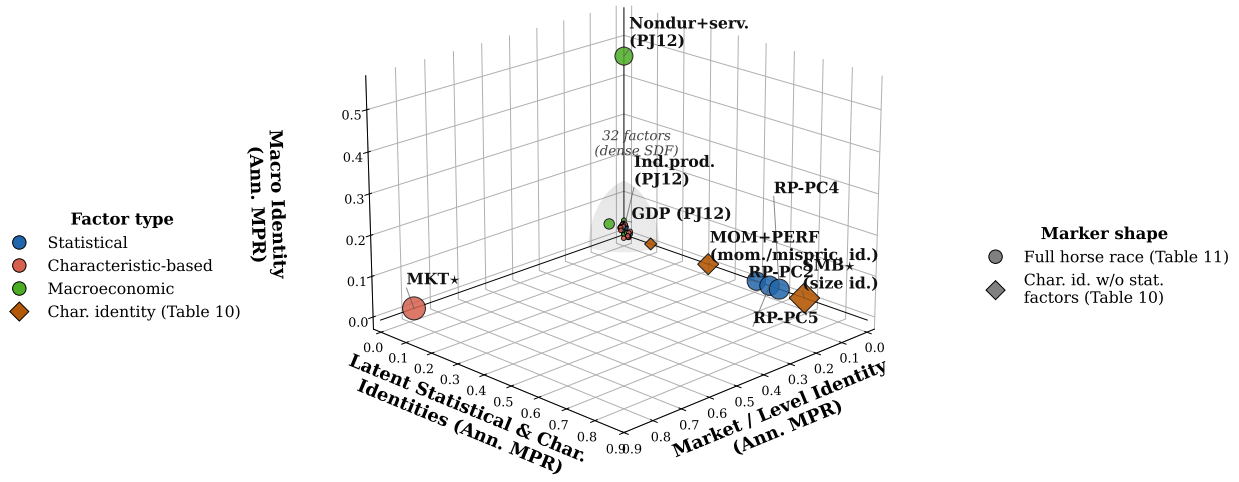

Figure 1: The Factor Zoo in Identity Space

The three axes correspond to the three main pricing identities: market/level (horizontal), latent statistical (depth, collapsing the three RP-PC identities), and macroeconomic (vertical). Each factor's coordinate equals its within-identity posterior probability times the annualised identity-level risk price. Marker size scales with the absolute unconditional posterior mean risk price. Circles are from the full horse race (Table 11); diamonds mark characteristic-based identities, size (SMB⋆) and momentum/mispricing (MOM, PERF), that emerge without statistical factors (Table 10) but are subsumed by PCs. The cluster near the origin contains the 32 factors that carry nonzero unconditional risk prices in the dense BMA-SDF (their positions are jittered proportionally to their unconditional risk prices).

identity, but both are fully subsumed by the latent statistical identity once risk-premia principal components (Lettau and Pelger, 2020) are included. The factor zoo is dense in animals, but its animals belong to just a handful of species. In geometric terms, our method recovers the asset pricing axes that span the cross-section of expected returns, identified not by return variance (as in standard factor analysis), but by cross-sectional pricing implications, and asks which observable factors best label each axis (Figure 1).

But how do we discover a factor identity? Two factors share an identity when they are near-substitutes in terms of their pricing implications: their covariances with asset returns are sufficiently similar that the data cannot determine which one drives the cross-sectional variation in expected returns. This near-substitutability is intimately connected to two well-known identification challenges in the literature: *weak factors*, whose covariances with asset

[1]E.g., Kozak et al. (2020); Bryzgalova et al. (2023); Dickerson et al. (2026); Kelly et al. (2019).

[2]This definition is simply an application of Leibniz's substitutivity principle to asset pricing: identical things are interchangeable *salva veritate*, where 'truth' here is measured in terms of implied risk premia.

returns are close to zero (Kan and Zhang, 1999a,b; Kleibergen, 2009), and *level factors*, whose covariances exhibit little variation across assets (Kleibergen and Zhan, 2020; Bryzgalova, Huang, and Julliard, 2023).[3] We show that for two factors capturing the same unique pricing component, their risk exposure estimates are not perfectly correlated only in two cases: either one of them is weak, or both are effectively level factors. Recognising these three complementary identification challenges, near-substitutability, weakness, and level-factor bias, we propose a Bayesian method that solves all three simultaneously.

Our method achieves two objectives that are missing from the existing literature. It detects and groups together factors that are close substitutes, based on the similarity of their pricing implications after controlling for all other factor exposures. By assigning near-substitute factors into a common identity, we average out noise in their risk exposures and remove the arbitrary choice among near-equivalent proxies, yielding risk price estimates that are more stable, better identified, and correspond to the pricing component they actually capture. A by-product is a natural treatment of level factors: a factor grouped with the intercept is treated as a level factor, and the intercept no longer absorbs its risk price; rather, it stabilises the estimate by not over-interpreting the distinction between two observationally equivalent pricing effects.

Beyond grouping, our method delivers a second layer of inference: for each detected identity, it reports the posterior probability that the identity enters the SDF, and, conditional on that, the probability that each member factor is the single best proxy for the identity's priced component. This two-level structure is especially valuable in the presence of weak factors, which generate risk exposures that are small on average but noisy, so their risk prices are weakly identified and highly unstable. In our framework, weak factors receive very low probabilities of being the best proxy for any identity: including them in the SDF delivers little incremental econometric benefit, measured by the increase in the attainable Sharpe ratio, relative to the prior.

[3] A level factor can be thought of as a special case of substitutable factors, because it arises in the limiting situation where the factor's cross-sectional risk exposures are nearly the same for every asset. Then the factor-implied component of expected returns is simply a constant shift applied equally to all assets, which is empirically indistinguishable from the intercept in cross-sectional regressions. Thus, level factors have no distinct pricing implication and their risk prices are not identifiable as long as a cross-sectional intercept is allowed.

Our method reveals the true identity of several well-known factors. First, the market portfolio and the first principal component of realised returns behave as level factors: they help capture the overall level of risk premia but contribute little to explaining their cross-sectional variation. The same is true of the intermediary capital ratio factor of He et al. (2017). Second, we uncover common identities among characteristic-based factors. Several factors designed to capture similar dimensions of risk share a common pricing identity: the size factors of Fama and French (2015) and Hou et al. (2015) are grouped together, as are the investment factors from both models; and among technical-signal factors, momentum and reversal strategies cluster with common latent statistical factors.

Third, consistent with the predictions of investment-based asset pricing (Berk et al., 1999; Zhang, 2005) but previously undocumented in terms of cross-sectional substitutability, the value factor of Fama and French (2015) tells the same pricing story as both investment factors. This result is visible in the data: the estimated loadings on the investment factors of Fama and French (2015) and Hou et al. (2015) both correlate at approximately 0.95 with the loadings on the value factor, even though their time-series correlation is only around 0.67. Fourth, the two leading intermediary asset pricing factors, the broker-dealer leverage factor of Adrian et al. (2014) and the capital ratio factor of He et al. (2017), carry distinct identities, consistent with their opposite cyclical implications (Haddad and Muir, 2021; Santos and Veronesi, 2022).

Fifth, and strikingly, macroeconomic factors reveal their identity only at long horizons. At the quarterly frequency, the link between macro variables and expected returns appears weak. But when measured over twelve quarters, cumulative growth in GDP, industrial production, and nondurable-plus-services consumption emerges robustly as a single pricing identity, echoing Parker and Julliard (2005) and Bryzgalova et al. (2025a) for consumption and complementing Angeletos et al. (2020) and Bryzgalova et al. (2025b), who show that the same aggregate shock drives these fundamentals over two- to three-year horizons. In a comprehensive horse race among statistical, characteristic-based, and macroeconomic factors, the macro identity survives even after controlling for latent statistical factors, while characteristic-based factors are largely displaced. That is, latent factors render most characteristics redundant but cannot subsume the pricing information contained in macroeconomic

fundamentals, establishing a clear hierarchy: macro identities carry genuinely distinct information that statistical factors alone do not capture.

That said, redundancy is not irrelevance. When latent statistical factors are excluded from the horse race, additional characteristic-based identities emerge: a denoised size identity (Daniel et al., 2020) and a momentum/mispricing identity. These are fully subsumed once risk-premia PCs (Lettau and Pelger, 2020) enter the comparison: the pricing variation these characteristics capture is genuine, but better measured by latent factors. More broadly, anomaly factors of the same economic category tend to converge to a common identity, consistent with the view that each characteristic style proxies for a shared underlying risk.

The remainder of the paper is organized as follows. Below, we review the most closely related literature and our contribution to it. Section 1 develops the theoretical framework and estimation method. Section 2 provides simulation evidence in realistically small samples. Section 3 presents our empirical findings, and Section 4 concludes. Additional details, results, and proofs are reported in the Appendix and the Internet Appendix.

Closely related literature. Our paper contributes to several interconnected strands of the asset pricing, macro-finance, and financial econometrics literatures.

Factor proliferation and the structure of the SDF. A large literature seeks to determine which of the hundreds of proposed factors drive the cross-section of expected returns (Cochrane, 2011; Harvey et al., 2016). Kozak et al. (2018, 2020) show that no-near-arbitrage conditions force most SDF variance onto a few high-eigenvalue principal components, and that the SDF is dense in characteristics space but sparse in PC space; Haddad et al. (2020) extend the argument to the conditional SDF, showing that the optimal factor-timing portfolio equals the pricing kernel. Our paper shifts the question from "which factors?" or "how many PCs?" to "how many distinct pricing identities?" That a handful of dominant PCs suffice to summarize the cross-section is consistent with the few discrete identities we recover. Our contribution is to discretize and label the redundancy structure the PC-based literature identifies but leaves as continuous rotations, mapping each group to a recognizable economic channel.

Factor selection, identification, and robust inference. Identification failures severely dis-

tort asset pricing tests: useless factors are spuriously found priced (Kan and Zhang, 1999a,b; Kleibergen, 2009; Kleibergen and Zhan, 2015; Gospodinov et al., 2014, 2019), level factors pose additional challenges (Kleibergen and Zhan, 2020; Bryzgalova et al., 2023), and several methods target factor selection directly (Feng et al., 2020; Giglio and Xiu, 2021; He et al., 2023; Pukthuanthong et al., 2019). Our paper pursues a distinct problem: factors that induce nearly the same cross-sectional risk exposures, of which the level-factor problem of Kleibergen and Zhan (2020) is a special case. Unlike the methods above, which ask "does this factor matter?", we ask "which factors are interchangeable?" Unlike the sequential protocol of Pukthuanthong et al. (2019), our unified Bayesian framework simultaneously forms identity groups, ranks proxies within each group, and estimates risk prices.

Bayesian approaches and model uncertainty. Bayesian methods have a long history in asset pricing (Harvey and Zhou, 1990; Pástor and Stambaugh, 2000; Pástor, 2000; Barillas and Shanken, 2018, 2017; Chib et al., 2020; Avramov, 2002; Anderson and Cheng, 2016; Avramov et al., 2023). Most relevant to our work, Bryzgalova et al. (2023) propose Bayesian-SDF and BMA-SDF estimators that search over quadrillions of models with a prior linking identification strength to risk-price variance, and Dickerson et al. (2026) generalise this to heterogeneous asset classes. Our innovation is to decompose risk prices into identity-level components, each driven by a single factor selected from competing proxies, yielding two layers of inference: identity-level probabilities (does this pricing component exist?) and within-identity factor probabilities (which proxy best represents it?). Methodologically, the cross-sectional step builds on the sum of single effects (SuSiE) model of Wang et al. (2020), which decomposes a sparse coefficient vector into single-effect components. Our setting departs from SuSiE in that both the dependent variable (expected returns) and the regressors (factor covariances) are latent. Hence, we embed the SuSiE-type step in a hierarchical structure that draws from the posterior of these latent inputs and propagates their uncertainty into risk price estimates—a Bayesian analogue of the Shanken (1992) correction. As in Pástor and Stambaugh (2000), the prior on each identity's risk price maps into Sharpe ratio beliefs. The structure nests standard Bayesian factor selection as a special case; when several factors share an identity, it reveals their substitutability.

Macroeconomic factors and long-horizon risk. Macroeconomic variables typically show

weak pricing power at short horizons despite their theoretical centrality. Parker and Julliard (2005) and Jagannathan and Wang (2007) show that measuring macro risk over multiple quarters dramatically improves cross-sectional pricing; Bryzgalova et al. (2025a,b) attribute this to a common priced shock that propagates slowly through macro aggregates; and Angeletos et al. (2020) identify a single business-cycle shock as the dominant driver of comovement in output, consumption, and investment. Our paper shows these findings reflect a single pricing identity. The 12-quarter cumulative growth rates of GDP, industrial production, and nondurable-plus-services consumption form one identity with cross-sectional covariance exposures correlated at 89–94%. This macro identity survives a horse race with latent statistical factors, which displace characteristic-based factors but cannot subsume macroeconomic fundamentals.

Investment-based asset pricing and the value–investment nexus. A parallel theoretical tradition derives expected returns from the production side of the economy (Cochrane, 1991; Berk et al., 1999; Zhang, 2005): firms' book-to-market ratios and investment rates both reflect the same latent mix of assets in place and growth options, so that long-short portfolios formed on either dimension generate near-identical cross-sectional risk exposures. The q-factor model of Hou et al. (2015) operationalizes this by constructing investment and profitability factors that subsume many anomalies previously attributed to value. A key but previously untested implication is that value and investment factors should be *near-substitutes in cross-sectional pricing*: because both characteristics reflect the same latent state variable, the risk exposures they generate across assets should be nearly collinear, even when their time-series returns differ. Our finding that HML, CMA, and IA form a single pricing identity provides direct evidence for this prediction.

Intermediary asset pricing. He et al. (2017) and Adrian et al. (2014) propose competing intermediary factors based on, respectively, market and book leverage of financial intermediaries. Haddad and Muir (2021) investigate whether these factors reflect intermediary-specific risk or economy-wide risk aversion. Our method provides a complementary lens: rather than testing the channel, we ask whether the factors are observationally equivalent in cross-sectional pricing. They are not. The HKM factor groups with the intercept as a level factor, whereas AEM forms a distinct identity, consistent with their opposite cyclical implications.

This also bears on the hedging of risk factors: as Herskovic et al. (2019) show, managing factor exposure requires knowing which factors represent distinct risk dimensions. Factor identities provide exactly this information: hedging within an identity is redundant; hedging across identities is where diversification value lies.

1 Theory and Method

This section develops the theoretical framework and estimation method. We begin with the identification challenge that motivates our approach. Subsection 1.1 uses a simple two-factor example to establish a key result (Proposition 1): even when two candidate factors have modest time-series correlation, their cross-sectional risk exposures can be nearly indistinguishable, making it impossible to reliably attribute the risk premium to one factor rather than the other. Moreover, using the same example, we show that whenever the estimated exposures are *not* perfectly collinear, one of two well-documented complications must be present: either one factor is weak, with covariances with test assets that are too small (Kan and Zhang, 1999a,b; Kleibergen, 2009), or both are effectively level factors, to which the asset loadings are too similar to each other (Kleibergen and Zhan, 2020; Bryzgalova et al., 2023).

This trio of identification challenges—near-substitute, weak, and level factors—renders traditional inference methods unreliable at best, and motivates the robust method we develop in the remainder of the section. Subsection 1.2 introduces a Bayesian single-identity approach that assigns posterior probabilities to each candidate factor based on their ability to explain the cross-section of expected returns. Subsection 1.3 extends this building block to multiple pricing identities, each capturing a distinct component of expected returns, and integrates time-series sampling uncertainty into the inference.

1.1 An Illustrative Example

Consider a one-factor linear SDF

$$M_t = 1 - \lambda^\star f_t, \tag{1}$$

where the priced factor $f_t \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$. The scalar $\lambda^\star$ represents the factor's price of risk. Since we work with *excess returns*, we normalize the mean of the SDF (and of the factor) to

one (and zero), without loss of generality. For brevity, we will refer to excess returns simply as "returns." The return on asset i is

$$r_{i,t} = c_i(\lambda^\star + f_t) + e_{i,t}, \qquad i = 1, \ldots, n, \tag{2}$$

where $e_{i,t} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_e^2)$ is the idiosyncratic component, independent of the systematic component f_t. The coefficient $c_i = \text{Cov}[r_{i,t}, f_t]$ is the covariance between asset i's return and the factor, which (because the factor has unit variance) also corresponds to the standard beta exposure of asset i. We treat these exposures as an inherent feature of the supply side of the economy, and assume $c_i \overset{\text{iid}}{\sim} F_c$, which is independent of all other stochastic parts in the example. We denote by

$$\mu_c := \mathbb{E}_i[c_i] = \int x \, \mathrm{d}F_c(x) \qquad \text{and} \qquad V_c := \text{Var}_i[c_i] = \int (x - \mu_c)^2 \, \mathrm{d}F_c(x) \tag{3}$$

the *cross-sectional* mean and variance of the risk exposures, respectively. We assume that the cross-sectional mean exposure $\mu_c \neq 0$; that is, the factor is priced on average across assets. Under the SDF specification in equation (1) and the fundamental asset pricing equation $\mathbb{E}[M_t \cdot r_{i,t}] = 0$, we must have

$$\mu_i := \mathbb{E}[r_{i,t}] = c_i \lambda^\star. \tag{4}$$

The return dynamics in equation (2) satisfy this condition by construction.

Now suppose that, in empirical work, a noisy proxy for f_t is also considered, defined as

$$\tilde{f}_t = \rho f_t + \sqrt{1 - \rho^2}\tilde{e}_t, \tag{5}$$

where $\tilde{e}_t \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$ and is independent of both f_t and $e_{i,t}$. That is, $\tilde{e}_t$ is a pure measurement error. Denote by $\tilde{c}_i = \text{Cov}(r_{i,t}, \tilde{f}_t)$ the exposure of asset i to $\tilde{f}_t$. Standard asset pricing tests evaluate how well these covariances explain the cross-section of expected returns by estimating the following regression:

$$\mu_i = \lambda_0 + c_i\lambda + \tilde{c}_i\tilde{\lambda} + \alpha_i, \qquad i = 1, \ldots, n, \tag{6}$$

where λ_0 is the mean pricing error, λ and $\tilde{\lambda}$ are the risk prices of f_t and $\tilde{f}_t$, respectively, and α_i is the (demeaned) pricing error for asset i. Under the ideal no-arbitrage benchmark, the expected returns must satisfy (4). We would expect $\lambda_0 = \tilde{\lambda} = 0$ and $\lambda = \lambda^\star$ (the true risk price) in (6).

In the regression (6), the independent variables c_i and $\tilde{c}_i$ must be perfectly correlated whenever $\rho \neq 0$, since $\tilde{c}_i = \rho c_i$. This holds regardless of how small $|\rho|$ is. Economically, factors that load on the same systematic risk may exhibit only low time-series correlation, yet their cross-sectional risk exposures across assets can be nearly indistinguishable.

Of course, the statement above holds only at the population level. Empirical work also requires estimating these risk exposures from the time-series observations $\left\{\{r_{i,t}\}_{i=1}^n, f_t, \tilde{f}_t\right\}_{t=1}^T$. For simplicity, we assume that the factors have been demeaned in sample. Then, we can estimate the two covariance terms as:

$$\hat{c}_i = \frac{1}{T}\sum_{t=1}^{T} r_{i,t} f_t, \qquad \hat{\tilde{c}}_i = \frac{1}{T}\sum_{t=1}^{T} r_{i,t}\tilde{f}_t.$$

Turning to the sample behaviours, we have the following result, the proof of which is in Appendix C.1.

Proposition 1. *In the example above, suppose the cross-sectional variance of the true factor exposures (V_c in equation (3)) and the correlation between $\tilde{f}_t$ and the true factor (ρ in equation (5)) are both non-zero constants as $T \to \infty$. Then the cross-sectional correlation between $\hat{c}_i$ and $\hat{\tilde{c}}_i$ satisfies* $\mathrm{Corr}_i(\hat{c}_i, \hat{\tilde{c}}_i) = \mathrm{sgn}(\rho) + O_p(T^{-1/2})$.

According to this proposition, once we rule out the two well-known issues in the cross-sectional asset pricing literature—namely, the presence of weak ($\rho \to 0$) or level ($V_c \to 0$) factors—empiricists must face a different kind of trouble. That is, the independent variables in equation (6) become almost perfectly collinear, as $\mathrm{sgn}(\rho) = \pm 1$. As a result, standard cross-sectional regressions cannot reliably separate the effects of the two factors, making it difficult to identify which one truly drives the variation in expected returns.

In the next subsection, we introduce an approach that is well suited to address this problem. Instead of delivering a single point estimate for each risk price, our approach quantifies the probability that each factor makes a distinct contribution to the SDF in any

observed return sample. More importantly, it explicitly accommodates the potentially strong correlations among factor exposures (e.g., c_i and $\tilde{c}_i$ in equation (6)) and groups together those factors that are statistically indistinguishable from one another in terms of cross-sectional pricing implications: that is, the factors with a common (pricing) identity. This is what most existing Bayesian (and frequentist) methods fail to achieve. As a preview, it is worth noting that our proposed method also handles two well-known complications in asset pricing tests: it can place a level factor into the same category as the common intercept, thereby properly identifying its risk price, and can eliminate weak factors with high probability.

1.2 The Single-Identity Case

We begin with the single-identity case: a single pricing component drives the cross-section of expected returns, and the data are used to learn which factor best represents it. Beyond building intuition, this case is the essential building block of the general multi-identity approach developed in Subsection 1.3.

Consider the following cross-sectional regression for mean asset returns:

$$\boldsymbol{\mu}_r = \boldsymbol{C}\boldsymbol{\lambda} + \boldsymbol{\alpha}, \tag{7}$$

where $\boldsymbol{\mu}_r = [\mu_1, \ldots, \mu_n]^\top$, with $\mu_i = \mathbb{E}[r_{i,t}]$, concatenates all the unconditional expected returns; $\boldsymbol{C} = [\mathbf{1}_n, \boldsymbol{c}_1, \ldots, \boldsymbol{c}_p]$ is an $n \times (p+1)$ matrix in which the first column is a vector of ones, and its $(j+1)$-th column ($j = 1, \ldots, p$), $\boldsymbol{c}_j$, equals $\mathrm{Cov}(\boldsymbol{r}_t, f_{j,t})$, which is an $n \times 1$ vector of covariances between asset returns and the j-th factor; $\boldsymbol{\alpha} \in \mathbb{R}^n$ is the vector of cross-sectional pricing errors; and $\boldsymbol{\lambda} = [\lambda_0, \lambda_1, \ldots, \lambda_p]^\top$ contains the risk prices associated with the factors, except that λ_0 captures the mean pricing error. As usual, we assume that all factors are demeaned and standardised. This extends equation (6) by allowing the covariances to be defined with respect to multiple (instead of two) factors.

We allow for model misspecification in equation (7). Pricing errors can arise due to, e.g., measurement errors, financial frictions, market inefficiencies, or the statistical limits to arbitrage. We assume that the cross-sectional pricing errors $\{\alpha_i\}_{i=1}^n$ are i.i.d. normal with zero mean and variance $\sigma_\alpha^2 > 0$, i.e., $\alpha_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\alpha^2)$, and are cross-sectionally independent of the covariance exposures $\boldsymbol{C}$. This form of model misspecification has been extensively

studied in the literature (Kan et al., 2013; Gospodinov et al., 2014; Giglio and Xiu, 2021; Da et al., 2025).[4]

Recall that our objective is to make valid inferences about the risk prices even when some of the covariance vectors are highly similar. To accomplish this, we take a Bayesian approach and assign priors to the risk prices $\boldsymbol{\lambda}$ in equation (7). Although prior choices are inherently subjective (and no priors can satisfy everyone) we aim to specify priors that (i) induce valid inferences when some columns in $\boldsymbol{C}$ are collinear, or when level and weak factors are present, as discussed in Proposition 1, (ii) yield analytically tractable posteriors, and (iii) incorporate economically meaningful restrictions. We introduce the priors and discuss how these criteria are met below.

Specifically, with the hyperparameters $\boldsymbol{\pi} = [\pi_0, \pi_1, \ldots, \pi_p]^\top$ and σ_0, we assume

$$\boldsymbol{\lambda} = \boldsymbol{\gamma} \cdot \lambda, \qquad \boldsymbol{\gamma} \sim \text{Mult}(1, \boldsymbol{\pi}), \qquad \lambda \sim \mathcal{N}(0, \sigma_0^2). \tag{8}$$

Here, the binary vector $\boldsymbol{\gamma} = [\gamma_0, \gamma_1, \ldots, \gamma_p]^\top$ with $\gamma_j \in \{0, 1\}$, drawn from the $(p+1)$-dimensional multinomial distribution, is such that $\sum_{j=0}^{p} \gamma_j = 1$, and has exactly one non-zero entry.[5] That is, the vector $\boldsymbol{\gamma}$ selects the factor that best represents the unique identity that prices assets, and the market price of risk of this factor identity is the scalar λ. If $\gamma_j = 1$ $(j \neq 0)$, the j-th factor is the only one that enters the SDF with a non-zero risk price λ drawn from $\mathcal{N}(0, \sigma_0^2)$. If $\gamma_0 = 1$, then all factors are excluded from the SDF and only the intercept remains. In general, the posterior means of $\{\gamma_j\}_{j=0}^{p}$ will be different from zero and one, expressing the probability that each factor is the best proxy for the identity.

Economic priors. The probability vector $\boldsymbol{\pi}$ reflects the prior beliefs assigned to each factor being the best representative of the single-identity. In our empirical analysis, we take

[4] An alternative specification, $\boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, \kappa\boldsymbol{\Sigma}_\alpha)$, where $\boldsymbol{\Sigma}_\alpha$ is the residual covariance matrix, has been studied by Pástor and Stambaugh (2000); Pástor (2000); Barillas and Shanken (2018); Bryzgalova et al. (2023). This formulation requires n/T to be small. Because our cross-section exceeds 100 assets and some exercises use quarterly data, we adopt the more parsimonious $\alpha_i \sim \mathcal{N}(0, \sigma_\alpha^2)$, which is also common in the literature.

[5] In other words, out of the p factors (plus the intercept), our prior postulates that only one of them (or none, if $\pi_0 = 1$, in which case only the intercept is selected) can enter the SDF as best proxy for the single-identity. This prior effectively induces "discrete choices" among factors. The same prior specification has been proposed in statistical genetics (Servin and Stephens, 2007). Our full methodology, described in the next section, relaxes this restriction.

an agnostic view by letting $\pi_0 = \pi_1 = \ldots = \pi_p = 1/(1+p)$. Nevertheless, stronger beliefs regarding certain factors can be directly incorporated into the elements $\pi_1, \ldots, \pi_p$ of this vector.

The prior volatility σ_0 of the risk price λ controls the variance of the SDF. The linear factor SDF associated with the cross-sectional regression in equation (7) is

$$M_t = 1 - \boldsymbol{\lambda}^\top \boldsymbol{f}_t = 1 - \lambda(\boldsymbol{\gamma}^\top \boldsymbol{f}_t), \tag{9}$$

where $\boldsymbol{f}_t = [f_{1,t}, \ldots, f_{p,t}]^\top$, λ is the price of risk of the single-identity, and we set $f_{0,t} \equiv 0$ so that selecting the intercept ($\gamma_0 = 1$) yields $M_t = 1$. As the factors have zero mean and unit variance, and $\mathbb{E}[M_t] = 1$, it follows that

$$\mathrm{Var}[M_t] = \mathbb{E}[\mathbb{E}[(M_t - 1)^2 \mid \boldsymbol{\gamma}]] = \sum_{j=1}^{p} \pi_j \, \mathbb{E}[(\lambda f_j)^2 \mid \gamma_j = 1] + \pi_0 \times 0 = (1 - \pi_0)\sigma_0^2.$$

It is common in the asset pricing literature to follow Hansen and Jagannathan (1991) (or relatedly, Cochrane and Saa-Requejo (2000)) in connecting the variance of the SDF to the maximal Sharpe ratio achievable in the economy (Kozak et al., 2020; Bryzgalova et al., 2023; Huang and Shi, 2025). Under the prior specification in equation (8), the parameter σ_0 captures the beliefs about the attractiveness of investment opportunities from the perspective of a mean-variance investor—that is, it represents a prior belief about the Sharpe ratio achievable with the single factor identity.

Summing up, the economic prior beliefs determine the hyperparameters σ_0 and $\boldsymbol{\pi}$, which are treated as known from now on. In most empirical applications we consider prior values of σ_0 corresponding to an annualized Sharpe ratio of 0.30 to 0.50 for the single factor identity.

Closed-form posteriors. Conditional on observing the average returns $\boldsymbol{\mu}_r$, the covariance matrix $\boldsymbol{C} = [\boldsymbol{c}_0, \boldsymbol{c}_1, \ldots, \boldsymbol{c}_p]$ with $\boldsymbol{c}_0 = \mathbf{1}_n$, and the variance of pricing errors σ_α^2, the posterior distribution of $(\lambda, \boldsymbol{\gamma})$ is given by the following proposition, the proof of which follows directly from the derivations in Appendix A.2.1.

Proposition 2. *Under the prior specification in equation (8) for the risk prices, we have*

the following posterior: for $j = 0, 1, \ldots, p$,

$$\lambda \,|_{\gamma_j=1,\,\boldsymbol{\mu}_r,\,\boldsymbol{C},\,\sigma_\alpha^2} \sim \mathcal{N}\left(m_j,\, s_j^2\right), \quad q_j := \Pr\left[\gamma_j = 1 \mid \boldsymbol{\mu}_r,\, \boldsymbol{C},\, \sigma_\alpha^2\right] = \frac{w_j}{\sum_{k=0}^{p} w_k},$$

where

$$m_j = \frac{\sigma_0^2 \boldsymbol{\mu}_r^\top \boldsymbol{c}_j}{\sigma_0^2 \|\boldsymbol{c}_j\|^2 + \sigma_\alpha^2}, \qquad s_j^2 = \frac{\sigma_0^2 \sigma_\alpha^2}{\sigma_0^2 \|\boldsymbol{c}_j\|^2 + \sigma_\alpha^2}, \qquad w_j = \pi_j \sqrt{\frac{s_j^2}{\sigma_0^2}} \exp\left(\frac{m_j^2}{2 s_j^2}\right)$$

The above posteriors can be further understood by considering the following univariate regression

$$\boldsymbol{\mu}_r = \lambda_j \, \boldsymbol{c}_j + \boldsymbol{\alpha}, \qquad \boldsymbol{\alpha} \sim \mathcal{N}(0, \sigma_\alpha^2 \mathbf{1}_n). \tag{10}$$

In this regression, let

$$\hat{\lambda}_j = \frac{\boldsymbol{c}_j^\top \boldsymbol{\mu}_r}{\|\boldsymbol{c}_j\|^2}, \qquad \sigma(\hat{\lambda}_j) = \frac{\sigma_\alpha}{\|\boldsymbol{c}_j\|}, \qquad z_j = \frac{\hat{\lambda}_j}{\sigma(\hat{\lambda}_j)}, \qquad \kappa_j = \frac{\sigma_0^2 \|\boldsymbol{c}_j\|^2}{\sigma_0^2 \|\boldsymbol{c}_j\|^2 + \sigma_\alpha^2}. \tag{11}$$

Then, $\hat{\lambda}_j$ is the OLS estimate of the slope coefficient; $\sigma(\hat{\lambda}_j)$ is its standard deviation, taking the variance of pricing errors σ_α^2 as known; z_j is the z-score for testing $H_0 : \lambda_j = 0$ against the alternative; κ_j is a signal-to-noise ratio measured under the prior $\lambda_j \sim \mathcal{N}(0,\, \sigma_0^2)$, which is the proportion of cross-sectional variation in risk premia explained by the covariance with factor j. Under these definitions, one can verify that in Proposition 2,

$$m_j = \kappa_j \hat{\lambda}_j, \qquad s_j^2 = \kappa_j \sigma^2(\hat{\lambda}_j), \qquad w_j = \pi_j \sqrt{1 - \kappa_j} \exp\left(\frac{\kappa_j}{2} z_j^2\right). \tag{12}$$

The posterior probability for the j-th factor entering the SDF is a monotone transformation of the absolute z-score $|z_j|$. When the factor is included in the SDF, the posterior mean and variance of its risk price are essentially the corresponding OLS estimates scaled by the shrinkage factor κ_j. If the j-th factor explains the cross-section of asset returns perfectly, hence $\sigma_\alpha^2 = 0$, we have $\kappa_j = 1$, and m_j is simply the OLS estimate of λ_j in equation (10). More generally, as $\sigma_\alpha^2 \to 0$, the exponential term in equation (12) dominates, and the method assigns increasingly higher posterior probability to the factors that deliver the

smallest pricing errors in the cross-sectional regression. This reflects a natural duality: σ_α^2 captures the overall pricing ability of the identity (a larger λ^2 implies a higher SDF variance and hence smaller residual mispricing variance) while z_j measures how well each individual factor serves as that identity's best representative.[6]

For ease of exposition in later sections, we introduce a function, $\mathcal{S}$ (for single-identity), that returns the posterior distributions defined in Proposition 2. Since this posterior distribution is determined by $\{\boldsymbol{\mu}_r, \boldsymbol{C}, \sigma_\alpha^2\}$ (as well as σ_0^2 and $\boldsymbol{\pi}$, which are pinned down by the economic prior beliefs), we write

$$\left\{\boldsymbol{m}, \boldsymbol{s}^2, \boldsymbol{q}, \theta\right\} \leftarrow \mathcal{S}\left(\boldsymbol{\mu}_r, \boldsymbol{C}, \sigma_\alpha^2\right), \tag{13}$$

where $\boldsymbol{m} = [m_0, \ldots, m_p]^\top$ is the vector of estimated risk prices (i.e. posterior means), $\boldsymbol{s}^2 = [s_0^2, \ldots, s_p^2]^\top$ is the vector of posterior variances (both conditional on being selected as the identity's representative) and $\boldsymbol{q} = [q_0, \ldots, q_p]^\top$ is the vector of posterior selection probabilities. That is, q_j is the posterior probability that the j-th factor is the best proxy for the single pricing identity.

Posterior properties. The single-identity approach summarised above can accommodate the econometric problem illustrated in Section 1.1. Consider the regression (6), which effectively introduces two collinear columns $\boldsymbol{c}_j = [c_1, \ldots, c_n]^\top$ and $\tilde{\boldsymbol{c}}_j = [\tilde{c}_1, \ldots, \tilde{c}_n]^\top$ such that $\tilde{\boldsymbol{c}}_j = \rho \boldsymbol{c}_j$. With some straightforward calculation, it can be shown that if the risk premia indeed satisfy equation (4), then

$$m_j = \lambda^\star + O_p(n^{-1/2}), \qquad \frac{m_{\tilde{j}}}{m_j} = \frac{1}{\rho} + O_p(n^{-1}). \tag{14}$$

Thus, the posterior mean in Proposition 2 is a consistent estimator of the true factor's risk price. Furthermore, the noisy proxy $\tilde{f}_t$ (as defined in equation (5)) will also be assigned

[6] This interpretation connects to the insight of Barillas and Shanken (2017), who show that for traded factor models, model comparison reduces to comparing excluded-factor alphas: how well each model prices the factors it omits. Our approach pursues a related objective at the identity level: within each identity, the method selects the factor that minimises pricing errors across a large cross-section of test assets. The analogy is tightest when the candidate factors are themselves included among the test assets, as is the case in several of our empirical specifications, since minimising cross-sectional pricing errors then subsumes pricing the excluded factors.

an equivalent risk price estimate, up to the rescaling coefficient $\rho = \|\tilde{\boldsymbol{c}}_j\|/\|\boldsymbol{c}_j\|$, which compensates for the smaller norm of the covariance vector. This is reasonable because these two factors have essentially the same cross-sectional covariances with the test assets; they differ only by scale. Because the two factors have almost identical explanatory power for the cross-section of mean returns, our approach assigns them non-zero posterior probabilities, q_j and $q_{\tilde{j}}$, which are characterised in the following proposition, the proof of which is given in Appendix C.2.

Proposition 3. *In the example given in Section 1.1, when the true expected returns are* $\boldsymbol{\mu}_r = \lambda^\star \boldsymbol{c}_j$, *for two sets of covariances* $\boldsymbol{c}_j$ *and* $\tilde{\boldsymbol{c}}_j$ *such that* $\tilde{\boldsymbol{c}}_j = \rho \boldsymbol{c}_j$, *the posterior probabilities calculated according to Proposition 2 are such that*

$$\frac{q_j}{q_{\tilde{j}}} \longrightarrow \frac{\pi_j}{\pi_{\tilde{j}}} \times |\rho| \times \exp\left[\frac{1}{2}\left(\frac{1}{\rho^2} - 1\right)\left(\frac{\lambda^\star}{\sigma_0}\right)^2\right] \tag{15}$$

with probability one, as $n \to \infty$. *If the true risk price* $\lambda^\star$ *is no less than the prior belief about the maximal Sharpe ratio* σ_0, *i.e.,* $\lambda^* \geq \sigma_0$, *and* $\pi_j = \pi_{\tilde{j}}$, *then with probability one* $q_j > q_{\tilde{j}}$.

Proposition 3 implies that, as long as the true risk price (also a measure closely related to the maximal Sharpe ratio achievable in the data) supports some improvement in investment opportunities relative to our prior beliefs, the method will always assign higher probability to the true factor.

If the true factor is a so-called level factor, then its covariances exhibit almost no cross-sectional variation. In other words, V_c, as defined in equation (3), is close to zero. As a result, $\boldsymbol{c}_j \approx \mu_c \mathbf{1}_n = \mu_c \boldsymbol{c}_0$. Thus, the level factor problem is essentially the same collinearity problem discussed above and can also be handled by our approach. However, the correlation between $\boldsymbol{c}_j$ and $\boldsymbol{c}_0$ is not well defined, since $\boldsymbol{c}_0$ has zero cross-sectional variance.

If the factor $\tilde{f}_t$ is a weak factor, i.e., $\rho = 0$ in equation (5), it no longer conveys the same pricing information as the true factor. In this case, the (hyper)parameter σ_0^2, which incorporates the empiricists' belief about the maximal Sharpe ratio, is key to addressing the problem. Specifically, under the posterior defined in Proposition 2, we can show that, if

$\sigma_0^2 < \infty$, then as the number of test assets $n \to \infty$, with probability one,

$$q_j \to 1 \quad \text{and} \quad q_{\tilde{j}} \to 0. \tag{16}$$

This can be seen directly by letting $\rho \to 0$ in (15) of Proposition 3 and enforcing $q_j + q_{\tilde{j}} = 1$. The weak factor will receive asymptotically zero posterior probability. In contrast, if we do not incorporate a prior belief about the achievable Sharpe ratio, i.e. if we consider (as in the frequentist setting) a flat prior (achievable by letting $\sigma_0^2 \to \infty$ in equation (8)), the weak factor will have a posterior probability of one of entering the SDF, crowding out the true factor. That is, as pointed out in Bryzgalova et al. (2023), the fragility of canonical factor selection in the presence of weak factors is just another manifestation of the well-known pitfall of model selection based on flat priors, the so-called Lindley–Bartlett paradox (Lindley, 1957; Bartlett, 1957).

Finally, we also define

$$\theta = \frac{\sum_{j=0}^{p} w_j}{1 + \sum_{j=0}^{p} w_j}, \tag{17}$$

where w_j is defined in Proposition 2. The scalar θ is the posterior probability that the data favour including at least one factor (or the intercept) over the null that expected returns are driven entirely by pricing errors; that is, $\theta = \Pr[\boldsymbol{\mu}_r \neq \boldsymbol{\alpha} \mid \text{data}]$, and $\boldsymbol{\mu}_r = \boldsymbol{\alpha}$ plays the role of the null 'model' in the construction of the Bayes factors on which the posterior probability θ is based.

1.3 The General Multi-Identity Approach

The single-identity approach introduced earlier is restrictive because, in reality, multiple distinct pricing identities—each reflecting a different source of systematic risk—are expected to contribute to the SDF. We now present a general approach that accommodates multiple pricing identities, each represented by a group of near-substitute factors.

In addition to handling multiple pricing identities, another key innovation here is that we also account for estimation uncertainties in $\boldsymbol{\mu}_r$ and $\boldsymbol{C}$. This is a distinguishing feature of the cross-sectional regressions in asset pricing: both the dependent (expected returns) and the independent (asset covariances with factors) variables are unknown and must be esti-

mated from the time-series data. Our Bayesian approach explicitly accounts for estimation uncertainty in the time-series, for expected returns and covariances, and in the cross-section, for risk prices. To proceed, a distributional assumption for returns and factors is needed. Concatenating the returns $\boldsymbol{r}_t \in \mathbb{R}^n$ and factors $\boldsymbol{f}_t \in \mathbb{R}^p$ into an $n' = n+p$ dimensional vector, our full time-series model is then specified as follows, for time $t = 1, \ldots, T$,

$$\boldsymbol{y}_t = \begin{pmatrix} \boldsymbol{r}_t \\ \boldsymbol{f}_t \end{pmatrix} \overset{\text{iid}}{\sim} \mathcal{N}\left(\boldsymbol{\mu},\, \boldsymbol{\Sigma}\right), \text{[7]} \quad \begin{pmatrix} \boldsymbol{r}_t \\ \boldsymbol{f}_t \end{pmatrix} \overset{\text{iid}}{\sim} \mathcal{N}\left(\boldsymbol{\mu},\, \boldsymbol{\Sigma}\right), \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_r \\ \boldsymbol{\mu}_f \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{rr} & \boldsymbol{\Sigma}_{rf} \\ \boldsymbol{\Sigma}_{rf}^\top & \boldsymbol{\Sigma}_{ff} \end{pmatrix}. \tag{18}$$

The cross-sectional pricing restriction requires that, as in equation (7),

$$\boldsymbol{\mu}_r = \boldsymbol{C}\boldsymbol{\lambda} + \boldsymbol{\alpha}, \qquad \boldsymbol{C} = [\mathbf{1}_n, \boldsymbol{\Sigma}_{rf}], \qquad \boldsymbol{\alpha} \sim \mathcal{N}(0, \mathbf{1}_n \sigma_\alpha^2), \tag{19}$$

where the standardisation of factors to unit variance ensures that $\boldsymbol{\Sigma}_{rf}$ coincides with the covariance exposures in $\boldsymbol{C}$.

We first demonstrate that we can approximate the posterior distribution of time-series moments $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ without considering the cross-sectional pricing equation (19). Note that the terms in the posterior involving $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are proportional to

$$\underbrace{-\frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2T}\sum_{t=1}^{T}(\boldsymbol{y}_t - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_t - \boldsymbol{\mu})}_{\text{time-series}} \underbrace{-\frac{1}{2T\sigma_\alpha^2}\left\|\boldsymbol{\mu}_r - \boldsymbol{C}\boldsymbol{\lambda}\right\|^2}_{\text{cross-section}} \underbrace{+\frac{1}{T}\log \pi(\boldsymbol{\mu}, \boldsymbol{\Sigma})}_{\text{prior}},$$

where $\pi(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the prior distribution of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ specified in Section 1.3.2.

The cross-sectional pricing restriction interacts with the time-series realized return distributions because $\boldsymbol{\mu}_r$ and $\boldsymbol{C}$ are both effectively functions (via subsetting) of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Unfortunately, this connection breaks conjugacy in the posterior of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, preventing closed-form updates of the involved parameters.

We therefore omit the cross-sectional component when updating $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Intuitively, the time-series likelihood provides the primary information about the first and second moments,

[7] When factors are persistent, as with the 12-quarter cumulative macro growth rates, we relax the iid assumption by replacing the posterior covariance with its Newey–West (Newey and West, 1987) counterpart, following Müller (2013). Details are in Appendix B.2.

reliably identifying $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ (and thus $\boldsymbol{\mu}_r$ and $\boldsymbol{C}$) even without the cross-sectional pricing restriction in equation (19). Empirically, the cross-sectional component provides only weak and indirect information about $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ compared to the dominant contribution of the time-series likelihood. This is because, according to standard distribution theory for quadratic forms, we have

$$\frac{1}{T\sigma_\alpha^2}\left\|\boldsymbol{\mu}_r - \boldsymbol{C}\boldsymbol{\lambda}\right\|^2 \sim O_p\left(\frac{n}{T}\right), \qquad \frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{y}_t - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_t - \boldsymbol{\mu}) \sim O_p(n).$$

This indicates that the magnitude of the cross-sectional term is negligible relative to the time-series term, being at least two to three orders of magnitude smaller (in our empirical applications, the latter term is 200 to 500 times larger than the former).

We decompose the full posterior distribution of $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \sigma_\alpha^2)$ into two parts:

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \sigma_\alpha^2 \mid \{\boldsymbol{y}_t\}_{t=1}^T) = p(\boldsymbol{\lambda}, \sigma_\alpha^2 \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \{\boldsymbol{y}_t\}_{t=1}^T) \times p(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \{\boldsymbol{y}_t\}_{t=1}^T). \tag{20}$$

We first present how to approximate the posterior in the cross-sectional step, $p(\boldsymbol{\lambda}, \sigma_\alpha^2 \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \{\boldsymbol{y}_t\}_{t=1}^T) \equiv p(\boldsymbol{\lambda}, \sigma_\alpha^2 \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ (by sufficiency of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for the cross-sectional pricing parameters), via variational inference in Section 1.3.1. Next, in Section 1.3.2, we introduce a factor-based prior used to discipline the estimation of $p(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \{\boldsymbol{y}_t\}_{t=1}^T)$.

1.3.1 Posterior inference for cross-sectional pricing

Conventional Bayesian approaches usually report the probability that each factor is selected. Our method, in comparison, is particularly useful in identifying multiple pricing *identities* supported by the data. Each identity captures a distinct component of the cross-section of expected returns. Multiple factors may proxy for the same component, and grouping them into a single identity acknowledges a key reality: the data often cannot distinguish among alternative measures of the same economic force.

To enable the selection of multiple pricing identities entering the SDF, we extend the

prior specification in equation (8), as follows:

$$\boldsymbol{\lambda} = \sum_{\ell=1}^{L} \boldsymbol{\lambda}_\ell, \quad \boldsymbol{\lambda}_\ell = \boldsymbol{\gamma}_\ell \cdot \lambda_\ell, \quad \boldsymbol{\gamma}_\ell \overset{\text{iid}}{\sim} \text{Mult}(1, \boldsymbol{\pi}), \quad \lambda_\ell \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_0^2), \quad \text{and} \quad \pi(\sigma_\alpha^2) \propto \frac{1}{\sigma_\alpha^2}, \tag{21}$$

encoding the SDF as a sum over the factor identities: $M_t = 1 - \sum_{\ell=1}^{L} \lambda_\ell \boldsymbol{\gamma}_\ell^\top \boldsymbol{f}_t$, where each $\boldsymbol{\gamma}_\ell$ selects a single factor (the "best proxy") to represent identity ℓ. When $L = 1$, the prior coincides with the one introduced in equation (8). This multi-identity prior is adapted from the sum of single effects (SuSiE) model (Wang, Sarkar, Carbonetto, and Stephens, 2020). The vector of risk prices, $\boldsymbol{\lambda}$, is decomposed into the sum of L single-identity components, $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_L$. Each component ℓ captures a candidate *pricing identity*: a distinct channel through which factors may drive expected returns. The scalar λ_ℓ is the risk price associated with identity ℓ, and the posterior probability θ_ℓ (defined in equation 17) quantifies how likely this identity is to belong to the SDF.

Within each identity ℓ, the selection vector $\boldsymbol{\gamma}_\ell$ picks a single "best representative" among the p candidate factors (and the intercept). Its posterior expectation, $\boldsymbol{q}_\ell$, is a probability distribution over these candidates: $q_{\ell,j}$ measures how likely factor j is to be the best single proxy for identity ℓ. Factors that receive high probabilities within the same identity are interpreted as near-substitutes: they share similar, if not identical, pricing implications, and the data cannot conclusively single out one over the others.

Each single-identity component has the same prior volatility σ_0, calibrated to correspond to a prior annualised Sharpe ratio between 0.3 and 0.5, as discussed in Section 1.2. Throughout our empirical analysis, we set $\boldsymbol{\pi} = 1/(1+p)\mathbf{1}_{p+1}$. That is, we assign equal prior probability to each factor being the best representative of any given identity.

According to equation (21), if we define

$$\boldsymbol{\mu}_{r,\ell} = \boldsymbol{\mu}_r - \sum_{\ell' \neq \ell} \boldsymbol{C}\boldsymbol{\lambda}_{\ell'} \equiv \boldsymbol{C}\boldsymbol{\lambda}_\ell + \boldsymbol{\alpha}, \tag{22}$$

what we have above is effectively the single-identity model discussed in Section 1.2. The left-hand side variable is the expected return residual when the ℓ-th identity is excluded. The posterior of $(\lambda_\ell, \boldsymbol{\gamma}_\ell)$ is given by Proposition 2. To be more specific, treating the residualised

expected returns $\boldsymbol{\mu}_{r,\ell}$, as well as the covariances in $\boldsymbol{C}$ and the variance of pricing errors σ_α^2, as known, we can define the posteriors of $(\lambda_\ell,\, \boldsymbol{\gamma}_\ell)$ for any $\ell = 1, \ldots, L$ and all $j = 0, 1, \ldots, p$,

$$\lambda_\ell \,|_{\gamma_{\ell,j}=1,\, \boldsymbol{\mu}_r,\, \boldsymbol{C},\, \boldsymbol{\lambda}_{-\ell},\, \sigma_\alpha^2} \sim \mathcal{N}\left(m_{\ell,j}, s_{\ell,j}^2\right), \quad q_{\ell,j} := \Pr[\gamma_{\ell,j} = 1 \mid \boldsymbol{\mu}_r,\, \boldsymbol{C},\, \boldsymbol{\lambda}_{-\ell},\, \sigma_\alpha^2], \tag{23}$$

where $\boldsymbol{\lambda}_{-\ell}$ denotes the collection of $\{\boldsymbol{\lambda}_{\ell'}\}_{\ell' \neq \ell}$. By definition, $\boldsymbol{\mu}_{r,\ell}$ in equation (22) is a sufficient statistic for $\boldsymbol{\mu}_r$ and $\boldsymbol{\lambda}_{-\ell}$ in deriving the conditional distributions in equation (23). According to Proposition 2,

$$\left\{\boldsymbol{m}_\ell,\, \boldsymbol{s}_\ell^2,\, \boldsymbol{q}_\ell,\, \theta_\ell\right\} \;\leftarrow\; \mathcal{S}\left(\boldsymbol{\mu}_{r,\ell},\, \boldsymbol{C},\, \sigma_\alpha^2\right). \tag{24}$$

The operator $\mathcal{S}(\cdot)$ is introduced in equation (13) which implements Proposition 2. The three vectors are defined as $\boldsymbol{m}_\ell = [m_{\ell,0},\, m_{\ell,1},\, \ldots,\, m_{\ell,p}]^\top$, $\boldsymbol{s}_\ell^2 = [s_{\ell,0}^2,\, s_{\ell,1}^2,\, \ldots,\, s_{\ell,p}^2]^\top$ and $\boldsymbol{q}_\ell = [q_{\ell,0},\, q_{\ell,1},\, \ldots,\, q_{\ell,p}]^\top$. The scalar θ_ℓ defined in equation (17) is the posterior probability that the ℓ-th identity as a whole contains at least one useful factor—which may even behave similarly to an intercept—conditional on the composition of other identities.

With these posteriors at hand, we can estimate $\boldsymbol{\lambda}_\ell$ by the posterior mean

$$\bar{\boldsymbol{\lambda}}_\ell = \mathbb{E}[(\lambda_\ell \cdot \boldsymbol{\gamma}_\ell) \mid \boldsymbol{\mu}_r,\, \boldsymbol{C},\, \boldsymbol{\lambda}_{-\ell},\, \sigma_\alpha^2] = \boldsymbol{m}_\ell \odot \boldsymbol{q}_\ell, \tag{25}$$

where $\odot$ denotes element-wise multiplication of two vectors (or matrices) of the same size.

The observation above naturally motivates a posterior inference scheme. That is, we can iteratively estimate (or update), for each identity ℓ, the conditional means $\boldsymbol{m}_\ell$ and posterior probabilities $\boldsymbol{q}_\ell$, by plugging the current risk price estimates $\bar{\boldsymbol{\lambda}}_\ell$ defined in equation (25) into the residualised expected returns in equation (22), until convergence. Points (i)–(iii) of Algorithm 1, Step 2(a), summarise this procedure.

Wang et al. (2020) draw an insightful analogy between algorithms of the sum of single effects (SuSiE) type, such as Algorithm 1 above, and the textbook *forward selection* procedure in regression analysis. Forward selection typically begins by choosing the single best independent variable through univariate regressions, for example, based on the t-statistics of the slope coefficients. It then iteratively adds, at each step, the variable whose inclusion most improves model fit, which is commonly done by regressing the current residuals (after

controlling for previously selected variables) on the remaining candidates and selecting the one with the strongest explanatory power.

Our approach follows a similar iterative logic but departs in a crucial way. Rather than selecting a single "best" variable at each step, it returns a *probability distribution* over all variables, informed by univariate regressions (see the discussion following Proposition 2). These distributions, namely $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_L$ as defined in equation (25), explicitly quantify the uncertainty about which variables (in our setting, the factor covariances) should be selected at each iteration.

Each such distribution naturally defines a *pricing identity*, say $\boldsymbol{q}_1$, consisting of asset pricing factors that behave similarly after controlling for other identities $\boldsymbol{q}_2, \ldots, \boldsymbol{q}_L$. When several factors within the same identity receive non-negligible probabilities, this suggests that they are capturing similar underlying economic forces that drive the cross-section of expected returns, rather than representing distinct effects.

For example, suppose the pricing identity $\ell = 1$ assigns non-negligible probabilities to two profitability factors, which are long-short portfolios of stocks sorted by operating profitability (Fama and French, 2015, FF) or by net income (Hou et al., 2015, HXZ): $q_{1,\text{FF}} = 0.5$, $q_{1,\text{HXZ}} = 0.5$. This pattern indicates that, after controlling for other factors, the two profitability factors convey largely the same pricing implications, and the observed return data cannot conclusively determine which one dominates. The two factors should therefore be interpreted as alternative measures of the same underlying profitability effect, rather than distinct sources of expected-return variation. Conversely, if sorting by net income produces a profitability factor that has a high probability of belonging to another useful pricing identity, we are likely to trace out a genuinely new economic mechanism.

Now, suppose that, in identity $\ell = 2$, the probabilities satisfy $q_{2,0} = 0.2, q_{2,j} = 0.5, q_{2,j'} = 0.3$, and all others are zero. In this scenario, factors j and j' are both useful and *at least* one of them should enter the SDF. However, because they are also grouped together with the intercept term—which itself commands a non-trivial probability of $q_{2,0} = 0.2$—these two factors collectively behave like a level factor. In other words, the cross-sectional variation in their exposures across assets is too small for the data to distinguish them from the intercept (in terms of explaining variation in $\boldsymbol{\mu}_r$). As a preview of some of our empirical results (which

may reinforce some readers' priors) the first principal components of most return panels, as well as the market portfolio, often exhibit precisely this type of behaviour.

The variance of the pricing errors is also estimated. In Appendix A.2.2, we show that, conditional on $\boldsymbol{\mu}_r$ and $\boldsymbol{C}$, we can estimate σ_α^2 as

$$\hat{\sigma}_\alpha^2 = \frac{1}{n}\left\{\|\boldsymbol{\mu}_r - \boldsymbol{C}\bar{\boldsymbol{\lambda}}\|^2 + \text{trace}\left(\boldsymbol{C}^\top \boldsymbol{C}\boldsymbol{V}_\lambda\right)\right\},$$

where $\bar{\boldsymbol{\lambda}} = \sum_{\ell=1}^{L} \boldsymbol{m}_\ell \odot \boldsymbol{q}_\ell$, $\boldsymbol{V}_\lambda = \sum_{\ell=1}^{L}\left[\text{diag}\{(\boldsymbol{s}_\ell^2 + \boldsymbol{m}_\ell^2) \odot \boldsymbol{q}_\ell\} - \bar{\boldsymbol{\lambda}}_\ell \bar{\boldsymbol{\lambda}}_\ell^\top\right]$, following the derivations in Appendix A.2.1 and A.2.2. This estimate can be incorporated into the algorithm for making inferences about the general model, as reflected in Step 2.c of Algorithm 1.

Remark 1. *Algorithm 1 computes a mean-field variational approximation (VA) to the true posterior distribution.*[8] *That is, we aim to use the following distribution*

$$\nu(\{\lambda_l, \boldsymbol{\gamma}_l\}_{\ell=1}^{L}, \sigma_\alpha^2 \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{\ell=1}^{L}\prod_{j=0}^{p} \nu(\lambda_l, \gamma_{\ell,j} = 1 \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \times \nu(\sigma_\alpha^2 \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

to approximate the true posterior distribution $p(\{\lambda_l, \boldsymbol{\gamma}_l\}_{\ell=1}^{L}, \sigma_\alpha^2 \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$. *The goal of VA is to find* $\nu(\cdot)$ *by minimizing the Kullback-Leibler divergence from* ν *to* p*, given the functional form of* ν*. We provide a formal justification for Algorithm 1 using VA in Appendix A.*

Variational inference plays an important role in our estimation by circumventing the "label-switching problem" that commonly arises in mixture models (Stephens, 2000). Because the likelihood is invariant to permutations of the identity labels, for any permutation $\mathcal{T} : \{1, \ldots, L\} \to \{1, \ldots, L\}$, we have

$$p\big(\{\lambda_\ell, \boldsymbol{\gamma}_\ell\}_{\ell=1}^{L}, \sigma_\alpha^2 \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}\big) = p\big(\{\lambda_{\mathcal{T}(\ell)}, \boldsymbol{\gamma}_{\mathcal{T}(\ell)}\}_{\ell=1}^{L}, \sigma_\alpha^2 \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}\big).$$

This symmetry makes the true posterior multimodal: each permutation of the L identities constitutes a distinct mode with identical probability. If one were to sample from this posterior via Markov chain Monte Carlo (MCMC), successive draws could freely switch identity labels (e.g., what is identity $\ell = 1$ in one draw may become identity $\ell = 3$ in the

[8] See Blei et al. (2017) for a comprehensive review of variational inference methods.

Algorithm 1 (Posterior Inference for Cross-Sectional Pricing). Making inferences about the multi-identity cross-sectional asset pricing model (19)–(21), conditional on $\boldsymbol{\mu}_r$ and $\boldsymbol{C}$.

Input: $\boldsymbol{\mu}_r \in \mathbb{R}^n$, $\boldsymbol{C} \in \mathbb{R}^{n\times(p+1)}$
$\boldsymbol{\pi} \in \mathbb{R}^{(p+1)} \equiv [1/(1+p), \ldots, 1/(1+p)]^\top$ (by default)
the prior variance of risk prices σ_0^2
the number of identities L

Step 1. Initialize $\bar{\boldsymbol{\lambda}}_\ell$ for all $\ell = 1, \ldots, L$; set $\bar{\boldsymbol{\lambda}} = \sum_{\ell=1}^L \bar{\boldsymbol{\lambda}}_\ell$ and $\sigma_\alpha^2 = \left\|\boldsymbol{\mu}_r - \boldsymbol{C}\bar{\boldsymbol{\lambda}}\right\|^2 / n$.

Step 2. Iterate the following steps until convergence:

a. For $\ell = 1, \ldots, L$, based on the current values of $\{\bar{\boldsymbol{\lambda}}_1, \ldots, \bar{\boldsymbol{\lambda}}_L\}$ and σ_α^2
   i. update the residualised expected returns $\boldsymbol{\mu}_{r,\ell} \leftarrow \boldsymbol{\mu}_r - \sum_{\ell'\neq\ell} \boldsymbol{C}\bar{\boldsymbol{\lambda}}_{\ell'}$;
   ii. apply the single-identity operator $\mathcal{S}(\cdot)$ as defined in equation (13) according to Proposition 2:
   $$\left\{\boldsymbol{m}_\ell,\, \boldsymbol{s}_\ell^2,\, \boldsymbol{q}_\ell,\, \theta_\ell\right\} \leftarrow \mathcal{S}\left(\boldsymbol{\mu}_{r,\ell},\, \boldsymbol{C},\, \sigma_\alpha^2\right);$$
   iii. update the risk prices: $\bar{\boldsymbol{\lambda}}_\ell \leftarrow \boldsymbol{m}_\ell \odot \boldsymbol{q}_\ell$ as in equation (25).

b. Update the aggregate risk price vector $\bar{\boldsymbol{\lambda}} \leftarrow \sum_{\ell=1}^L \bar{\boldsymbol{\lambda}}_\ell$.

c. Update the pricing-error variance
$$\sigma_\alpha^2 \leftarrow \frac{1}{n}\left\{\left\|\boldsymbol{\mu}_r - \boldsymbol{C}\bar{\boldsymbol{\lambda}}\right\|^2 + \operatorname{trace}\left(\boldsymbol{C}^\top\boldsymbol{C}\boldsymbol{V}_\lambda\right)\right\},$$
where $\boldsymbol{V}_\lambda = \sum_{\ell=1}^L \left[\operatorname{diag}\{\boldsymbol{q}_\ell \odot (\boldsymbol{s}_\ell^2 + \boldsymbol{m}_\ell^2)\} - \bar{\boldsymbol{\lambda}}_\ell\bar{\boldsymbol{\lambda}}_\ell^\top\right]$ and $\boldsymbol{m}_\ell^2$ is the element-wise squares of $\boldsymbol{m}_\ell$.

Output. Report $\left\{\bar{\boldsymbol{\lambda}}_\ell,\, \boldsymbol{q}_\ell, \boldsymbol{m}_\ell, \boldsymbol{s}_\ell^2,\, \theta_\ell\right\}_{\ell=1}^L$ and σ_α^2

next) making it impossible to aggregate draws and characterize uncertainty for any given identity. Variational inference, as implemented in Algorithm 1, avoids this problem because it converges to a single mode of the posterior, effectively fixing the identity labeling throughout the estimation.

Another advantage of variational inference is its rapid convergence. In our implementation, the algorithm runs until the parameter estimates stabilize. Specifically, every 500 iterations we compute the maximum absolute change in parameter estimates relative to 500 iterations earlier, and terminate when this change falls below a pre-specified tolerance (set to 0.001).

1.3.2 Estimation uncertainty in the expected returns and covariances

In Algorithm 1, we estimate the risk prices and the cross-sectional variance of pricing errors, given the expected returns $\boldsymbol{\mu}_r$ and the factor loadings $\boldsymbol{C}$. For empiricists seeking to make valid inferences about risk prices, especially when quantifying uncertainty with confidence

or credible intervals, it is crucial to recognize that both $\boldsymbol{\mu}_r$ and $\boldsymbol{C}$ are estimated from the observed returns and factors. Intuitively, estimating $\boldsymbol{\mu}_r$ introduces sampling uncertainty into the pricing errors. Estimating $\boldsymbol{C}$, however, incorporates measurement errors in the covariates into the risk price estimates. This mechanism is closely related to the Shanken (1992) correction in the classic two-pass regression introduced by Fama and MacBeth (1973), which adjusts inference to account for first-stage estimation error in factor loadings.

A natural starting point is the invariant and noninformative prior, i.e., $\pi(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-\frac{n'+1}{2}}$ $(n' = n + p)$, which imposes minimal structure on the expected returns and risk exposures. As we show in Appendix B.1, this prior specification results in the standard normal-inverse-Wishart posterior distribution.

However, in high-dimensional settings, such as when there are many test assets but relatively short return histories, the benchmark prior can lead to near-singular posteriors for the risk exposures. This problem is particularly pronounced when quarterly observations of macroeconomic factors are used, as in our empirical studies. As a result, the posteriors of risk exposures can become extremely sensitive to noise (often substantial in large cross-sections of asset returns) that is present in the sample.

To address this issue, we consider a factor-based prior that explicitly imposes a low-rank factor structure on expected returns and covariances. Specifically, $\pi\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ is defined through the following model:

$$\boldsymbol{\mu} = \boldsymbol{B}\boldsymbol{v} + \boldsymbol{u}, \qquad \boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{V}\boldsymbol{B}^{\top} + \boldsymbol{U}, \tag{26}$$

where $\boldsymbol{u}$ is a vector of n' elements, $\boldsymbol{U} = \text{diag}\{\sigma^2_{u,1}, \ldots \sigma^2_{u,n}, \sigma^2_{u,n+1}, \ldots \sigma^2_{u,n+p}\}$ is an $n' \times n'$ diagonal matrix. $\boldsymbol{B}$ is an $n' \times d$ matrix with row vectors $\boldsymbol{\beta}_1^{\top}, \ldots, \boldsymbol{\beta}_{n'}^{\top}$. $\boldsymbol{v}$ is a $d \times 1$ vector, and $\boldsymbol{V}$ is a $d \times d$ matrix. We are mostly interested in the case where $d \ll n'$. The priors for $(\boldsymbol{u}, \boldsymbol{B}, \boldsymbol{U})$ and $(\boldsymbol{v}, \boldsymbol{V})$ are specified as

$$\pi(\boldsymbol{u}, \boldsymbol{B}, \boldsymbol{U}) \propto |\boldsymbol{U}|^{-\frac{d+1}{2}}, \quad \pi(\boldsymbol{v}, \boldsymbol{V}) \propto |\boldsymbol{V}|^{-\frac{d+1}{2}}. \tag{27}$$

The factor-based parameterisation of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ introduces a factor structure into expected returns and the covariance matrix. This approach significantly reduces the dimensionality—

from $O(n'^2)$ free parameters to $O(n'd + d^2)$ free parameters—and helps stabilise posterior inference when the return time-series is short. Because the triplet $(\boldsymbol{B}, \boldsymbol{v}, \boldsymbol{V})$ is shared across all test assets (and factors), their estimation can borrow strength from information along the cross-sectional dimension, which is particularly useful in settings with many test assets.[9]

Under the factor-based prior defined in equations (26)–(27), we sample $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ using a Gibbs sampler, with the detailed steps summarized in Proposition B.1. Appendix B further provides the detailed derivation of the posterior distribution. In empirical applications, we always use a large value of d (specifically, $d = 40$, which still satisfies $d \ll n'$ in our applications) to ensure that the common component of $\boldsymbol{\Sigma}$, i.e., $\boldsymbol{BVB}^\top$, can adequately capture the factor covariances. In other words, we do not impose a low-dimensional representation of $\boldsymbol{\Sigma}$ ex ante, as is often done in principal component analysis.

Our full Bayesian approach is hierarchical. As equation (20) suggests, we first draw posterior samples of $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \mid \{\boldsymbol{r}_t, \boldsymbol{f}_t\}_{t=1}^T$, where $(\boldsymbol{\mu}_r, \boldsymbol{C})$ are part of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then, for each draw of $(\boldsymbol{\mu}_r, \boldsymbol{C})$, we use Algorithm 1 to recompute the conditional risk price estimates $\{\boldsymbol{m}_\ell\}_{\ell=1}^L$, their selection-weighted means $\{\bar{\boldsymbol{\lambda}}_\ell\}_{\ell=1}^L$, and the posterior identity probabilities $\{\theta_\ell\}_{\ell=1}^L$. Algorithm 2 summarises the implementation. This approach can be viewed as a Bayesian analogue of the Shanken correction: it accounts for estimation errors in expected returns and risk exposures directly through their posterior distributions, rather than via an ex post variance adjustment.

We fix the within-identity posterior factor probabilities, $\{\boldsymbol{q}_\ell\}_{\ell=1}^L$, in executing Algorithm 2. Specifically, we first use Algorithm 1 to estimate $\{\boldsymbol{q}_\ell\}_{\ell=1}^L$ conditional on point estimates of the mean returns and factor covariance matrix, $(\widehat{\boldsymbol{\mu}}_{\boldsymbol{r}}, \widehat{\boldsymbol{C}})$. Next, for each draw of $(\boldsymbol{\mu}_r, \boldsymbol{C})$, we update all other parameters in the cross-sectional step, conditional on the fixed $\{\boldsymbol{q}_\ell\}_{\ell=1}^L$. This implementation ensures that identity memberships do not vary across different draws of $(\boldsymbol{\mu}_r, \boldsymbol{C})$. Otherwise, because of the label-switching problem emphasized in the previous section, we will encounter multimodality in the Bayesian updating, which makes it difficult to summarize the posterior distribution. Moreover, as the simulation studies in Section 2 suggest, using the point estimates of the mean returns and factor covariance matrix, $(\widehat{\boldsymbol{\mu}}_{\boldsymbol{r}}, \widehat{\boldsymbol{C}})$,

[9] See Pati et al. (2014) for Bayesian treatments of this type of prior specification. The modelling choice is also adopted by Fan et al. (2008) within a frequentist framework for large covariance matrix estimation.

Algorithm 2 (Estimation Uncertainty). Incorporating estimation uncertainty in the expected returns and risk exposures.

Input: $\{\boldsymbol{r}_t,\, \boldsymbol{f}_t\}_{t=1}^T$
$\boldsymbol{\pi} \in \mathbb{R}^{(p+1)} \equiv [1/(1+p), \ldots, 1/(1+p)]^\top$ (by default)
the prior variance of risk prices σ_0^2
the number of identities L
the Monte Carlo sample size S

Step 0. Use Algorithm 1 to estimate $\{\boldsymbol{q}_\ell\}_{\ell=1}^L$ conditional on point estimates of the mean returns and factor covariance matrix, $(\widehat{\boldsymbol{\mu}}_{\boldsymbol{r}}, \widehat{\boldsymbol{C}})$.

Step 1. For $s = 1, 2, \ldots, S$, draw samples from the posteriors of $(\boldsymbol{\mu}_r,\, \boldsymbol{C}) \mid \{\boldsymbol{r}_t,\, \boldsymbol{f}_t\}_{t=1}^T$ according to Proposition B.1 in Appendix B, and denote the sample by

$$(\boldsymbol{\mu}_r^{(1)},\, \boldsymbol{C}^{(1)}), \ldots, (\boldsymbol{\mu}_r^{(S)},\, \boldsymbol{C}^{(S)}).$$

Step 2. For each sample $(\boldsymbol{\mu}_r^{(s)},\, \boldsymbol{C}^{(s)})$ and conditional on $\{\boldsymbol{q}_\ell\}_{\ell=1}^L$ obtained in step 0, calculate $\{\bar{\boldsymbol{\lambda}}_\ell^{(s)}, \boldsymbol{m}_\ell^{(s)}, \theta_\ell^{(s)}\}$, $\ell = 1, \ldots, L$, according to Algorithm 1.

rather than their varying posterior draws, to estimate $\{\boldsymbol{q}_\ell\}_{\ell=1}^L$ helps recover the underlying true set of factors and exclude irrelevant factors with high probability.

2 Simulations

To investigate the finite-sample performance of our Bayesian estimator, we build a simple simulation setting that includes redundant priced factors, strong but unpriced factors, and weak factors that are entirely uncorrelated with asset returns.

We calibrate the simulation setting to mirror the empirical analysis in Section 3, using the same cross-section of 102 equity portfolios from January 1974 to December 2019, as detailed in Section 3.1. In particular, monthly excess returns $\boldsymbol{r}_t$ are assumed to be driven by four latent variables $\boldsymbol{v}_t = (v_{1t}, \ldots, v_{4t})^\top$, as follows:

$$\boldsymbol{r}_t = \widehat{\boldsymbol{\mu}}_{\boldsymbol{r}} + \widehat{\boldsymbol{\beta}}_{\boldsymbol{v}} \boldsymbol{v}_t + \boldsymbol{\epsilon}_t,\ \boldsymbol{v}_t \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I_4}),\ \boldsymbol{\epsilon}_t \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{0}, \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}),\ \text{and}\ \widehat{\boldsymbol{\mu}}_{\boldsymbol{r}} = \widehat{\text{Cov}}(\boldsymbol{r}_t, \boldsymbol{v}_t)\boldsymbol{\lambda}_{true} + \widehat{\boldsymbol{\alpha}}. \quad (28)$$

To calibrate $\widehat{\text{Cov}}(\boldsymbol{r}_t, \boldsymbol{v}_t)$, $\widehat{\boldsymbol{\beta}}_{\boldsymbol{v}}$, and $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$, we extract the first four PCs of asset returns, and regress $\boldsymbol{r}_t$ on them to obtain $\widehat{\boldsymbol{\beta}}_{\boldsymbol{v}}$, which coincide with $\widehat{\text{Cov}}(\boldsymbol{r}_t, \boldsymbol{v}_t)$ because the PCs are mutually uncorrelated. $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$ is calibrated as the sample covariance matrix of $\widehat{\boldsymbol{\epsilon}}_t$ from the same regression.[10] To investigate the identification of the level factor, we treat v_{1t} as a level

[10] The sample covariance of $\widehat{\boldsymbol{\epsilon}}_t$ is singular because PCs are linear transformations of $\boldsymbol{r}_t$. We therefore add

factor, that is, we impose that $\widehat{\text{Cov}}(\boldsymbol{r}_t, v_{1t}) = \widehat{c} \cdot \mathbf{1}_n$, where $\widehat{c}$ is calibrated as the average return loadings on PC1.

We further presume that $\widehat{\text{Cov}}(\boldsymbol{r}_t, \boldsymbol{v}_t)\boldsymbol{\lambda}_{true}$ partially explains the cross-sectional spreads of mean returns, with $\boldsymbol{\lambda}_{true} = (-0.12, 0.14, -0.17, 0)^\top$. The risk prices of the first three latent factors map into annualized Sharpe ratios of 0.4, 0.5, and 0.6. This calibration yields a cross-sectional R^2 of 52%, denoting a moderate level of model misspecification. $\widehat{\boldsymbol{\alpha}}$ collects pricing errors orthogonal to betas: $\widehat{\boldsymbol{\alpha}}^\top \widehat{\boldsymbol{\beta}}_{\boldsymbol{v}} = \mathbf{0}$. Finally, v_{4t} has a zero risk price, despite being a strong factor in explaining time-series variation in returns.

We construct two sets of redundant factors as noisy proxies for v_{1t} and v_{2t}:

$$f_{1t} = \frac{1}{\sqrt{2}} v_{1t} + \frac{1}{\sqrt{2}} \mathcal{N}(0,1), \quad f_{2t} = \frac{1}{\sqrt{2}} v_{1t} + \frac{1}{\sqrt{2}} \mathcal{N}(0,1),$$

$$f_{3t} = \sqrt{\frac{2}{3}} v_{2t} + \sqrt{\frac{1}{3}} \mathcal{N}(0,1), \quad f_{4t} = \sqrt{\frac{1}{2}} v_{2t} + \sqrt{\frac{1}{2}} \mathcal{N}(0,1).$$

The covariances of assets with f_{1t} and f_{2t} are constant across assets, that is, $\boldsymbol{c}_1 = \boldsymbol{c}_2 = \frac{\widehat{c}}{\sqrt{2}} \mathbf{1}_n$. Thus, $(\mathbf{1}_n, \boldsymbol{c}_1, \boldsymbol{c}_2)$ constitute an identity of level factors. Moreover, f_{3t} and f_{4t} are noisy proxies for v_{2t}, where f_{3t} has a larger signal-to-noise ratio than f_{4t}. This specification implies that $\boldsymbol{c}_3 = \text{Cov}(\boldsymbol{r}_t, f_{3t})$ and $\boldsymbol{c}_4 = \text{Cov}(\boldsymbol{r}_t, f_{4t})$ are proportional to $\widehat{\boldsymbol{\beta}}_{v,2}$, so $(\boldsymbol{c}_3, \boldsymbol{c}_4)$ form the second identity of redundant factors.

Finally, we simulate a single priced factor f_{5t}, a strong yet unpriced factor f_{6t}, and a weak factor f_{7t} as follows: $f_{5t} = v_{3t}$, $f_{6t} = v_{4t}$, and $f_{7t} \sim \mathcal{N}(0,1)$. A proper econometric framework should identify three pricing identities, $(\mathbf{1}_n, \boldsymbol{c}_1, \boldsymbol{c}_2)$, $(\boldsymbol{c}_3, \boldsymbol{c}_4)$, and $\boldsymbol{c}_5$, and exclude the unpriced covariance terms, $\boldsymbol{c}_6$ and $\boldsymbol{c}_7$.

We simulate a representative sample of 600 monthly observations and use it to exemplify whether our Bayesian inference, introduced in Section 1.3, successfully recovers the factor identities. Table 1 displays the estimation results based on $L = 8$ and the prior parameter σ_0 set to an annualized prior Sharpe ratio of 0.4. Our method assigns 90–100% posterior probability to the first three identities. The first identity comprises three level factors $(\mathbf{1}_n, \boldsymbol{c}_1, \boldsymbol{c}_2)$, with within-identity posterior factor probabilities of 22.5%, 34.8%, and 42.7%, respectively.

$0.001 \times \boldsymbol{I}_n$ to the sample covariance of $\widehat{\boldsymbol{\epsilon}}_t$ and use the resulting regularized matrix as $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$.

The second identity includes only $\boldsymbol{c}_5$, whereas the third identity comprises $\boldsymbol{c}_3$ and $\boldsymbol{c}_4$. The remaining five identities, referred to as unselected, have identical posterior probabilities of only 22.7%. Factor probabilities are also mechanically identical across these low-probability identities and are not informative about the cross-sectional pricing ability of factors. In our empirical application, we report only the selected identities with posterior probabilities above 50%.

Given the within-identity factor probabilities $\{\boldsymbol{q}_\ell\}_{\ell=1}^8$, we estimate the posterior distribution of risk prices, using the approach described in Section 1.3.2. We present both conditional and unconditional factor risk prices. In the column "$p(\boldsymbol{m}_\ell \mid \text{data})$", we report the factor risk prices conditional on the factors being selected. In particular, we report the posterior mean and 90% credible intervals of risk prices for factors with probability above 1% in at least one of the selected identities. For instance, we estimate risk prices for f_5 based on the posterior draws in identity 2 (column $\ell = 2$), not on those in identities 4–8 (columns $\ell = 4$ to 8), although the posterior probability of $\boldsymbol{c}_5$ is 10% in the latter five identities. Accordingly, conditioned on f_5 being selected in identity 2, the posterior mean of its risk price is -0.196, with a 90% credible interval of $(-0.263, -0.131)$. We also report the unconditional posterior mean of the factor risk prices in the column "$\mathbb{E}[\boldsymbol{\lambda} \mid \text{data}]$", defined as $\mathbb{E}[\sum_{\ell=1}^L \boldsymbol{m}_\ell \odot \boldsymbol{q}_\ell \mid \text{data}]$. The unconditional posterior mean of f_5's risk price is about -0.2, almost identical to its conditional mean. Lastly, in the row labeled "$\mathbb{E}[\lambda_\ell \mid \text{data}]$", we report the posterior mean of the identity-level risk price, defined as $\mathbb{E}[\sum_j q_{\ell,j}\, m_{\ell,j} \mid \text{data}]$, for the selected identities. For example, the risk price of the third pricing identity ($\ell = 3$) is 0.153.[11]

In contrast, the GMM estimator entirely mischaracterizes the factor identities. For instance, the weak factor (f_{7t}) has a t-statistic of -2.272, whereas it is excluded by our Bayesian method. Moreover, risk prices of factors within the same identity often have opposite signs due to collinearity in their covariance exposures to returns. For $\boldsymbol{c}_3$ and $\boldsymbol{c}_4$, GMM estimates of the risk prices are 0.287 and -0.149, neither of which is statistically significant.

Remark 2. *The three estimation challenges—namely level factors, collinearity, and weak identification—invalidate the traditional approaches because of the nearly singular matrix* $\boldsymbol{C}$.

[11] In computing $\mathbb{E}[\lambda_\ell \mid \text{data}]$, we impose a sign normalization. For each identity ℓ, we take as reference the covariance vector of the factor with the highest within-identity posterior probability, and flip the sign of other factors' risk prices whenever their covariance vectors are negatively correlated with that reference.

Table 1: Bayesian Inference in One Simulated Sample

	Bayesian estimates										GMM estimates		λ_{true}
	Factor prob. within identities, $\boldsymbol{q}_\ell$								$p(\boldsymbol{m}_\ell \mid \text{data})$	$\mathbb{E}[\boldsymbol{\lambda} \mid \text{data}]$			
	$\ell=1$	$\ell=2$	$\ell=3$	$\ell=4$	$\ell=5$	$\ell=6$	$\ell=7$	$\ell=8$	mean & 90% CI		$\hat{\lambda}_{gmm}$	t-stat.	
$\mathbf{1}_n$	0.225	0	0	0.018	0.018	0.018	0.018	0.018	0.045 (-0.018, 0.109)	0.010	-0.446	(-2.088)	0.103
$\boldsymbol{c}_1$	0.348	0	0	0.029	0.029	0.029	0.029	0.029	-0.071 (-0.171, 0.028)	-0.025	0.156	(0.536)	-0.163
$\boldsymbol{c}_2$	0.427	0	0	0.026	0.026	0.026	0.026	0.026	-0.065 (-0.158, 0.026)	-0.028	-0.867	(-3.188)	-0.163
$\boldsymbol{c}_3$	0	0	0.654	0.144	0.144	0.144	0.144	0.144	0.151 (0.078, 0.223)	0.112	0.287	(1.014)	0.177
$\boldsymbol{c}_4$	0	0	0.346	0.152	0.152	0.152	0.152	0.152	0.158 (0.080, 0.237)	0.068	-0.149	(-0.458)	0.204
$\boldsymbol{c}_5$	0	1.000	0	0.100	0.100	0.100	0.100	0.100	-0.196 (-0.263, -0.131)	-0.201	-0.145	(-2.313)	-0.173
$\boldsymbol{c}_6$	0	0	0	0.168	0.168	0.168	0.168	0.168		0.019	0.088	(1.300)	0
$\boldsymbol{c}_7$	0	0	0	0.363	0.363	0.363	0.363	0.363		-0.032	-0.700	(-2.272)	0
Ident. prob.	0.895	0.999	0.960	0.227	0.227	0.227	0.227	0.227					
$\mathbb{E}[\lambda_\ell \mid \text{data}]$	0.063	-0.196	0.153										

This table compares Bayesian and GMM estimates for a single representative simulated sample ($T = 600$). We use Algorithm 1, with $L = 8$ and $\sigma_0 = 0.4/\sqrt{12}$, to estimate the posterior factor probabilities within each of the eight identities ($\{\boldsymbol{q}_\ell\}_{\ell=1}^8$). Given the within-identity factor probabilities $\{\boldsymbol{q}_\ell\}_{\ell=1}^8$, we estimate the posterior distribution of risk prices, using the approach described in Section 1.3.2. We present both conditional and unconditional factor risk prices. In the column "$p(\boldsymbol{m}_\ell \mid \text{data})$", we report the factor risk prices conditional on the factors being selected (both the conditional posterior mean and 90% credible intervals). In the column "$\mathbb{E}[\boldsymbol{\lambda} \mid \text{data}]$", we display the unconditional posterior mean of the factor risk prices, defined as $\mathbb{E}[\sum_{\ell=1}^L \boldsymbol{m}_\ell \odot \boldsymbol{q}_\ell \mid \text{data}]$. The posterior identity probabilities are defined as the posterior means of $\{\theta_\ell\}_{\ell=1}^L$ computed across all draws of time-series parameters $(\boldsymbol{\mu}_{\boldsymbol{r}}, \boldsymbol{C})$. In the row labeled "$\mathbb{E}[\lambda_\ell \mid \text{data}]$", we report the posterior mean of the identity market price of risk, defined as $\mathbb{E}[\sum_j q_{\ell,j} m_{\ell,j} \mid \text{data}]$. For the GMM estimation, we use an identity weighting matrix and report the point estimates, with t-statistics in parentheses. The final column displays the true risk prices in the simulated model.

An alternative way to handle the singularity issue is to employ the singular value decomposition (SVD) of $\boldsymbol{C}$*, i.e.,* $\boldsymbol{C} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$*, where* $\boldsymbol{U}$ *and* $\boldsymbol{V}$ *are orthonormal, and* $\boldsymbol{D}$ *is diagonal. The expected excess returns can be rewritten as* $\boldsymbol{\mu}_{\boldsymbol{r}} = \boldsymbol{C}\boldsymbol{\lambda} = \boldsymbol{U}\boldsymbol{D} \cdot \boldsymbol{V}^\top\boldsymbol{\lambda}$*. One can use only the large components of* $\boldsymbol{C}$ *in cross-sectional pricing. For instance, in the same simulated example as in Table 1,* $\text{diag}(\boldsymbol{D}) = (14.00, 2.05, 1.83, 1.39, 0.15, 0.10, 0.08, 0.07)^\top$*, implying that there are only four large components. Therefore, one can use the first four columns of* $\boldsymbol{U}$ *to price asset returns, which helps avoid the identification issues originating from small singular values of* $\boldsymbol{C}$*.*

However, SVD does not always deliver transparent factor identities. Note that the first four rows of $\boldsymbol{V}^\top$ *correspond to the factor identities, which are shown below:*

$$\boldsymbol{V}^\top = \begin{array}{c} \begin{matrix} \mathbf{1}_n & \boldsymbol{c}_1 & \boldsymbol{c}_2 & \boldsymbol{c}_3 & \boldsymbol{c}_4 & \boldsymbol{c}_5 & \boldsymbol{c}_6 & \boldsymbol{c}_7 \end{matrix} \\ \begin{pmatrix} -0.72 & 0.46 & 0.52 & -0.01 & -0.03 & 0.00 & 0.03 & -0.02 \\ 0.01 & 0.01 & -0.02 & -0.56 & -0.53 & -0.64 & -0.01 & 0.00 \\ -0.05 & -0.01 & -0.02 & 0.51 & 0.39 & -0.76 & -0.07 & -0.02 \\ -0.05 & -0.04 & 0.01 & -0.03 & -0.03 & 0.06 & -0.99 & -0.06 \end{pmatrix} \end{array}$$

For example, the first grouped covariance equals $-0.72 \times \mathbf{1}_n + 0.46 \times \boldsymbol{c}_1 + 0.52 \times \boldsymbol{c}_2$*, successfully compressing three level factors. Although we detect clear grouping of* $(\mathbf{1}_n, \boldsymbol{c}_1, \boldsymbol{c}_2)$ *and of* $\boldsymbol{c}_6$*, the second and third rows consist of not only* $(\boldsymbol{c}_3, \boldsymbol{c}_4)$ *but also* $\boldsymbol{c}_5$*. Therefore, SVD in this example fails to recover the pricing identities of* f_3*,* f_4*, and* f_5*.*

Table 2 reports the empirical frequencies of factor retention in each of the five identities across 2,000 simulations of three different sample sizes ($T \in \{200, 600, 1000\}$). In each simulated path, we include both the common intercept $\mathbf{1}_n$ and the seven observable factors (f_{1t} to f_{7t}) within the Bayesian inference, as explained in Algorithm 1, by setting $L = 8$ and considering three prior parameters of risk prices, $\sigma_0 \in \{0.3/\sqrt{12}, 0.4/\sqrt{12}, 0.5/\sqrt{12}\}$, which map into identity-level prior Sharpe ratios of 0.3–0.5 per annum. For the singleton identities such as $\boldsymbol{c}_5$, $\boldsymbol{c}_6$, and $\boldsymbol{c}_7$, we determine if the posterior factor probability is greater than a certain threshold (60% to 95%) in any selected identity. For identities with more than one factor, we need to compute the sum of the factor probabilities within each identity. For example, in Table 1, the sum of posterior probabilities of $(\mathbf{1}_n, \boldsymbol{c}_1, \boldsymbol{c}_2)$ is 100% in a chosen identity (see column $\ell = 1$ with a 90% posterior identity probability), exceeding all the thresholds. Therefore, $(\mathbf{1}_n, \boldsymbol{c}_1, \boldsymbol{c}_2)$ is identified as a priced identity in that simulation.

According to Table 2, our Bayesian estimator correctly identifies the true pricing identities and eliminates the useless factors ($\boldsymbol{c}_6$ and $\boldsymbol{c}_7$) in most simulations. For $T = 200$, which roughly matches the number of quarterly observations in real data, our approach successfully selects $(\mathbf{1}_n, \boldsymbol{c}_1, \boldsymbol{c}_2)$, $(\boldsymbol{c}_3, \boldsymbol{c}_4)$, and $\boldsymbol{c}_5$ in 91.8%, 75.4%, and 87.6% of the simulations, when using $\sigma_0 = 0.4/\sqrt{12}$ in the estimation and employing a threshold of 60% in the selection. In finite samples, there is a clear tradeoff between the size and power of a test. When using a smaller prior variance σ_0 or a larger threshold in the test, we tend to reduce the probability of mischaracterizing the unpriced factor, at the cost of lower statistical power of detecting priced identities. In empirical analysis, we perform the estimation for three values of σ_0 considered in Table 2, to ensure that our economic conclusions are robust to different, economically reasonable, priors. The final column of Table 2 displays the average (within-identity) factor probabilities. For the factors with an identical signal-to-noise ratio, such as f_1 and f_2, their posterior factor probabilities are roughly the same. In contrast, since f_4 is noisier than f_3, its posterior probability is slightly lower than that of f_3, on average.

Table 2: The Frequency of Retaining Risk Factors in Simulations

		60%	65%	70%	75%	80%	85%	90%	95%	post. factor prob. (mean)
						Panel A. $\sigma_0 = 0.3/\sqrt{12}$				
T = 200	$(\mathbf{1}_n, \boldsymbol{c}_1, \boldsymbol{c}_2)$	0.924	0.924	0.923	0.922	0.922	0.921	0.919	0.917	0.551, 0.188, 0.188
	$(\boldsymbol{c}_3, \boldsymbol{c}_4)$	0.716	0.716	0.715	0.712	0.705	0.695	0.685	0.650	0.466, 0.268
	$\boldsymbol{c}_5$	0.866	0.865	0.864	0.862	0.859	0.849	0.836	0.820	0.866
	$\boldsymbol{c}_6$	0.112	0.110	0.108	0.106	0.099	0.091	0.083	0.067	0.143
	$\boldsymbol{c}_7$	0	0	0	0	0	0	0	0	0.064
T = 600	$(\mathbf{1}_n, \boldsymbol{c}_1, \boldsymbol{c}_2)$	0.990	0.990	0.990	0.990	0.990	0.990	0.990	0.989	0.540, 0.224, 0.226
	$(\boldsymbol{c}_3, \boldsymbol{c}_4)$	0.908	0.908	0.907	0.904	0.896	0.892	0.878	0.855	0.552, 0.356
	$\boldsymbol{c}_5$	0.981	0.979	0.978	0.977	0.975	0.972	0.969	0.961	0.978
	$\boldsymbol{c}_6$	0.016	0.016	0.015	0.013	0.012	0.010	0.007	0.006	0.053
	$\boldsymbol{c}_7$	0	0	0	0	0	0	0	0	0.077
T = 1,000	$(\mathbf{1}_n, \boldsymbol{c}_1, \boldsymbol{c}_2)$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.523, 0.237, 0.239
	$(\boldsymbol{c}_3, \boldsymbol{c}_4)$	0.965	0.965	0.964	0.963	0.961	0.954	0.945	0.926	0.571, 0.391
	$\boldsymbol{c}_5$	0.998	0.998	0.998	0.997	0.996	0.995	0.994	0.990	0.997
	$\boldsymbol{c}_6$	0.002	0.002	0.002	0.002	0.001	0	0	0	0.038
	$\boldsymbol{c}_7$	0	0	0	0	0	0	0	0	0.082
						Panel B. $\sigma_0 = 0.4/\sqrt{12}$				
T = 200	$(\mathbf{1}_n, \boldsymbol{c}_1, \boldsymbol{c}_2)$	0.918	0.918	0.917	0.916	0.916	0.914	0.911	0.907	0.470, 0.225, 0.226
	$(\boldsymbol{c}_3, \boldsymbol{c}_4)$	0.754	0.754	0.752	0.748	0.745	0.738	0.730	0.709	0.456, 0.309
	$\boldsymbol{c}_5$	0.876	0.875	0.872	0.868	0.863	0.856	0.847	0.832	0.873
	$\boldsymbol{c}_6$	0.124	0.122	0.119	0.116	0.112	0.102	0.096	0.080	0.149
	$\boldsymbol{c}_7$	0.001	0	0	0	0	0	0	0	0.065
T = 600	$(\mathbf{1}_n, \boldsymbol{c}_1, \boldsymbol{c}_2)$	0.987	0.987	0.987	0.987	0.987	0.987	0.987	0.987	0.429, 0.277, 0.281
	$(\boldsymbol{c}_3, \boldsymbol{c}_4)$	0.934	0.933	0.930	0.928	0.924	0.917	0.911	0.892	0.540, 0.391
	$\boldsymbol{c}_5$	0.982	0.982	0.981	0.979	0.977	0.974	0.970	0.966	0.981
	$\boldsymbol{c}_6$	0.024	0.021	0.020	0.018	0.016	0.015	0.011	0.007	0.052
	$\boldsymbol{c}_7$	0	0	0	0	0	0	0	0	0.076
T = 1,000	$(\mathbf{1}_n, \boldsymbol{c}_1, \boldsymbol{c}_2)$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.408, 0.295, 0.297
	$(\boldsymbol{c}_3, \boldsymbol{c}_4)$	0.976	0.976	0.976	0.976	0.974	0.969	0.965	0.949	0.551, 0.423
	$\boldsymbol{c}_5$	0.999	0.998	0.998	0.997	0.997	0.996	0.995	0.992	0.998
	$\boldsymbol{c}_6$	0.004	0.004	0.003	0.002	0.002	0.001	0	0	0.032
	$\boldsymbol{c}_7$	0	0	0	0	0	0	0	0	0.082
						Panel C. $\sigma_0 = 0.5/\sqrt{12}$				
T = 200	$(\mathbf{1}_n, \boldsymbol{c}_1, \boldsymbol{c}_2)$	0.916	0.916	0.915	0.913	0.910	0.909	0.906	0.902	0.421, 0.247, 0.250
	$(\boldsymbol{c}_3, \boldsymbol{c}_4)$	0.771	0.770	0.769	0.766	0.763	0.755	0.746	0.726	0.448, 0.330
	$\boldsymbol{c}_5$	0.877	0.876	0.875	0.871	0.866	0.860	0.849	0.837	0.875
	$\boldsymbol{c}_6$	0.132	0.129	0.126	0.124	0.121	0.112	0.102	0.087	0.153
	$\boldsymbol{c}_7$	0.005	0.004	0.004	0.003	0.002	0.001	0.001	0.001	0.065
T = 600	$(\mathbf{1}_n, \boldsymbol{c}_1, \boldsymbol{c}_2)$	0.987	0.987	0.987	0.987	0.987	0.987	0.986	0.984	0.372, 0.305, 0.310
	$(\boldsymbol{c}_3, \boldsymbol{c}_4)$	0.942	0.941	0.940	0.935	0.930	0.926	0.916	0.903	0.518, 0.420
	$\boldsymbol{c}_5$	0.984	0.982	0.981	0.980	0.978	0.974	0.971	0.966	0.981
	$\boldsymbol{c}_6$	0.025	0.023	0.022	0.018	0.018	0.016	0.012	0.007	0.049
	$\boldsymbol{c}_7$	0	0	0	0	0	0	0	0	0.072
T = 1,000	$(\mathbf{1}_n, \boldsymbol{c}_1, \boldsymbol{c}_2)$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.352, 0.323, 0.324
	$(\boldsymbol{c}_3, \boldsymbol{c}_4)$	0.981	0.980	0.979	0.978	0.977	0.976	0.971	0.962	0.528, 0.450
	$\boldsymbol{c}_5$	0.999	0.998	0.998	0.998	0.996	0.995	0.994	0.992	0.997
	$\boldsymbol{c}_6$	0.004	0.004	0.004	0.002	0.002	0.001	0	0	0.028
	$\boldsymbol{c}_7$	0	0	0	0	0	0	0	0	0.079

This table reports the empirical frequency of factor retention in each of the five identities across 2,000 simulations of three different sample sizes ($T \in \{200, 600, 1000\}$). In each simulated path, we include both the common intercept $\mathbf{1}_n$ and the seven observable factors (f_{1t}–f_{7t}) into the Bayesian inference, as explained in Algorithm 1, by setting $L = 8$ and considering three prior parameters of risk prices, $\sigma_0 \in \{0.3/\sqrt{12}, 0.4/\sqrt{12}, 0.5/\sqrt{12}\}$. For the singleton identity (e.g., $\boldsymbol{c}_5$, $\boldsymbol{c}_6$, and $\boldsymbol{c}_7$), we assess whether the posterior factor probability is greater than a certain threshold (60% to 95%) in any selected identity. For identities with more than one factor, we compute the sum of the factor probabilities within each identity. If the sum of posterior probabilities exceeds the threshold (60% to 95%), this identity will be selected by our Bayesian method. The final column reports the average (within-identity) factor probabilities.

Table 3: Within-Identity Factor Probabilities Across Different Prior Specifications

	$\sigma_0 = 0.3/\sqrt{12}$			$\sigma_0 = 0.4/\sqrt{12}$			$\sigma_0 = 0.5/\sqrt{12}$		
	$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 1$	$\ell = 2$	$\ell = 3$
$\mathbf{1}_n$	0.541	0	0	0.387	0	0	0.327	0	0
$\boldsymbol{c}_1$	0.079	0	0	0.085	0	0	0.082	0	0
$\boldsymbol{c}_2$	0.380	0	0	0.528	0	0	0.590	0	0
$\boldsymbol{c}_3$	0	0	0.016	0	0	0.002	0	0	0
$\boldsymbol{c}_4$	0	0	0.984	0	0	0.998	0	0	1.000
$\boldsymbol{c}_5$	0	1.000	0	0	1.000	0	0	1.000	0
$\boldsymbol{c}_6$	0	0	0	0	0	0	0	0	0
$\boldsymbol{c}_7$	0	0	0	0	0	0	0	0	0
Ident. prob.	0.971	0.977	0.933	0.966	0.978	0.959	0.963	0.977	0.966

This table presents the Bayesian estimates in a simulated sample ($T = 200$). We use Algorithm 1, with $L = 8$ and $\sigma_0 \in \{0.3/\sqrt{12}, 0.4/\sqrt{12}, 0.5/\sqrt{12}\}$, to estimate the posterior factor probabilities within each of the eight identities ($\{\boldsymbol{q}_\ell\}_{\ell=1}^8$). The final row reports the posterior identity probabilities ($\{\theta_\ell\}_{\ell=1}^8$). We display the identities with posterior probabilities above 50%.

True factors within the same identity are not always selected simultaneously. In Table 3, we report the factor probabilities for the selected identities in another simulated sample ($T = 200$). When $\sigma_0 = 0.4/\sqrt{12}$ or $0.5/\sqrt{12}$, although we still recover all three true priced identities, only $\boldsymbol{c}_4$ commands a large factor probability of nearly 100% in identity $\ell = 3$. Nevertheless, when $\sigma_0 = 0.3/\sqrt{12}$, our approach assigns a posterior probability of 1.6% to $\boldsymbol{c}_3$, categorizing $\boldsymbol{c}_3$ and $\boldsymbol{c}_4$ into the same identity. Two caveats are noteworthy. First, in finite samples, not all true factors are guaranteed to be selected, but our approach will pick at least one useful factor within the same identity with high probability. Second, as usual, it is important to assess robustness across several prior parameter values— something that we do thoroughly in our empirical analysis.

3 Empirical Analysis

We apply our Bayesian method to three types of factors proposed in the previous literature (Connor, 1995): macroeconomic, statistical, and characteristic-based. Subsection 3.1 describes the data. In Subsection 3.2, we compare small-scale factor models with shared economic foundations and test whether these seemingly similar factors are classified into a single identity or selected into distinct identities. We conduct more extensive comparisons among the three types of factors in Subsection 3.3. Subsection 3.4 further investigates the sparsity of the SDF. In the main text, we report the estimation results based on a prior identity-level Sharpe ratio of 0.4 per annum, with $\sigma_0 = 0.4/\sqrt{4}$ for quarterly data and $\sigma_0 = 0.4/\sqrt{12}$ for

monthly data. Additional results for two prior Sharpe ratios, 0.3 and 0.5, are reported in the Appendix. Moreover, as explained in the methodological section, an implicit assumption in our prior specification is that the SDF model is sparse at the identity level. We choose the maximum number of identities to be ten, that is, $L_{max} = 10$, following the recommendation in Wang et al. (2020).

3.1 Data

As test assets, we study a large cross-section of 102 equity portfolios from Serhiy Kozak's website. The same data have been studied in past literature (e.g., Haddad et al. (2020) and Giglio et al. (2024)). In particular, ten decile portfolios are constructed for each of the 51 firm characteristics. We include both the long and short legs (i.e., the decile 10 and decile 1 portfolios) in our analysis, resulting in 102 equity portfolio returns. As shown in Table A1, these anomalies can be categorized into seven types: value-related, investment, profitability, trading frictions, issuance, momentum & reversal, and other characteristics. Given that the portfolios sorted by "standardized unexpected earnings" are available starting in October 1973 and that we need quarterly return data for testing macro factors, the full sample runs from January 1974 to December 2019.

We focus on 31 observable (both tradable and nontradable) factors. As characteristic-based factors, we include 19 prominent pricing factors: the Fama-French five factors (Fama and French, 2015), the hedged returns on these five factors (Daniel et al., 2020), the momentum factor (Jegadeesh and Titman, 1993), the q-factors (Hou et al., 2015), the expected growth factor (Hou et al., 2021), two mispricing factors (Stambaugh and Yuan, 2017), and short- and long-horizon behavioral factors (Daniel et al., 2020). For aggregate macro factors, we consider quarterly growth rates of GDP, industrial production, (durable, nondurable, and service) consumption, total factor productivity (TFP), and hours worked. For the unemployment rate, we use its quarterly change as the factor. Finally, we include aggregate liquidity from Pástor and Stambaugh (2003) and two intermediary factors: the AEM (Adrian et al., 2014) factor based on book leverage of broker-dealers, and the HKM (He et al., 2017) factor using market leverage of a wide set of financial intermediaries. Details of the factors are described in Table A2 of the Appendix.

In empirical applications, all variables are standardized to have unit variance per period. That is, the covariances between test asset returns and factors are interpreted as correlations, and we compare asset pricing factors based on their ability to explain mean asset returns in Sharpe ratio units.

3.2 Factor Models with Shared Economic Foundations

We first examine the asset pricing factors that are designed to capture the same underlying sources of economic risk.

Intermediary asset pricing theories (e.g., He and Krishnamurthy (2012, 2013)) predict that the marginal utility of financial intermediaries is salient for cross-sectional pricing. Yet there is no consensus on how to measure their marginal utilities. We investigate the two most popular measures in this literature. Adrian et al. (2014) focus on broker-dealer firms and use the book equity and debt data reported in Flow of Funds accounts. The AEM intermediary factor is defined as the shock to broker-dealer firms' book leverage. In contrast, He et al. (2017) construct shocks to the market-equity capital ratio of primary dealers, dubbed the HKM intermediary factor. Given the closely related economic motivations yet rather different measurement methods, do these intermediary factors explain the cross-section of mean returns in a homogeneous or heterogeneous way?

We answer this question by comparing the pricing abilities of AEM and HKM factors, controlling for the market factor as in He et al. (2017). The sample spans from 1974Q1 to 2017Q3, the end date of the AEM factor. Table 4 shows that our Bayesian method selects two distinct pricing identities. The first identity consists of purely level factors ($\mathbf{1}_n$, MKT, HKM). Although, like the market factor, the HKM factor cannot explain the cross-sectional spreads of mean returns, its risk price is sizable: the conditional posterior mean is 0.25, with a 90% credible interval of $(0.05, 0.45)$. The second identity is composed of the AEM intermediary factor, which carries a significant risk price of 0.44 per quarter. The estimates are also robust to alternative priors, as confirmed in Table A3 in the Appendix.

Are the empirical findings in Table 4 consistent with the previous literature? The fact that both intermediary factors carry positive risk prices aligns with Adrian et al. (2014) and He et al. (2017). However, there is a reciprocal relationship between the capital ratio

(HKM) and leverage (AEM). In theory, if the two intermediary factors reflect the same financial distress in the intermediary sector, their risk prices should have opposite signs, contradicting empirical observations. According to our test, the AEM and HKM factors fall into distinct identities, suggesting that they capture rather heterogeneous dimensions of financial intermediaries, both of which are important components of the SDF.

The observation that the market portfolio functions as a level factor in cross-sectional pricing aligns with classic evidence that the security market line is too flat (Jenson et al., 1972; Fama and French, 1992). Our paper explains why estimating the market risk price is empirically difficult: asset returns tend to have very similar covariance exposures to the market portfolio, making its risk price hard to identify. Our method assigns the market excess return to the same identity as the common intercept $\mathbf{1}_n$ and other level factors, thereby reliably recovering the market risk price from the data.

For two level factors, e.g., f_{1t} and f_{2t}, the correlation between returns' exposures to them—$\text{Cov}(\boldsymbol{r}_t, f_{1t})$ and $\text{Cov}(\boldsymbol{r}_t, f_{2t})$—is entirely uninformative about their factor identity. For instance, although the time-series correlation between MKT and HKM is about 0.79, the cross-sectional correlation between their covariance loadings is only 0.37. The underlying reason is that the cross-sectional variance of $\text{Cov}(\boldsymbol{r}_t, f_t)$, for a level factor f_t, converges to zero as the time-series sample size increases. This observation aligns exactly with Proposition 1 explained in the methodological section.

Unlike macroeconomic factors that are based on equilibrium asset pricing models, statistical factor analysis, such as PCA, is motivated by the arbitrage pricing theory in Ross (1976): the leading PCs of asset returns also determine cross-sectional risk premia. Following the previous literature using PCA factors (Chamberlain and Rothschild, 1983; Connor and Korajczyk, 1986, 1988; Kozak et al., 2020), we extract the five largest PCs and use them to explain mean returns. One pitfall of canonical PCA is that it omits the information embedded in the first moment of returns; that is, returns' exposures to latent factors should explain the mean returns. Lettau and Pelger (2020) instead develop a new variant of PCA, dubbed risk-premia PCA (RP-PCA), to extract dominant latent factors that explain both time-series and cross-sectional variation in equity returns. We estimate the first five

Table 4: HKM vs. AEM Intermediary Factors

	Factor prob. within identities, $\boldsymbol{q}_\ell$		$p(\boldsymbol{m}_\ell \mid \text{data})$	$\mathbb{E}[\boldsymbol{\lambda} \mid \text{data}]$
	$\ell = 1$	$\ell = 2$	mean & 90% CI	
$\mathbf{1}_n$	0.591	0	0.177 (0.035, 0.317)	0.105
MKT	0.162	0	0.197 (0.038, 0.353)	0.032
HKM	0.247	0	0.248 (0.049, 0.446)	0.061
AEM	0	1.000	0.438 (0.167, 0.700)	0.438
Ident. prob.	0.980	0.952		
$\mathbb{E}[\lambda_\ell \mid \text{data}]$	0.198	0.438		

This table presents the Bayesian estimates of two intermediary factors, HKM (He et al., 2017) and AEM (Adrian et al., 2014), controlling for a common intercept ($\mathbf{1}_n$) and the market factor (MKT). The data are at the quarterly frequency, spanning from 1974Q1 to 2017Q3. In empirical applications, all variables are standardized to have unit variance per period. The data description is in Section 3.1. We use Algorithm 1, with $L = 4$ and $\sigma_0 = 0.4/\sqrt{4}$, to estimate the posterior factor probabilities within identities ($\{\boldsymbol{q}_\ell\}_{\ell=1}^L$). The second last row reports the posterior identity probabilities ($\{\theta_\ell\}_{\ell=1}^L$). We display the identities with posterior probabilities above 50%. Conditional on $\{\boldsymbol{q}_\ell\}_{\ell=1}^L$, we estimate the posterior distribution of factor risk prices using the approach described in Section 1.3.2. We present both conditional and unconditional factor risk prices. In the column "$p(\boldsymbol{m}_\ell \mid \text{data})$", we report the factor risk prices conditional on the factors being selected (both the conditional posterior mean and 90% credible intervals). In the column "$\mathbb{E}[\boldsymbol{\lambda} \mid \text{data}]$", we display the unconditional posterior mean of the factor risk prices, defined as $\mathbb{E}[\sum_{\ell=1}^L \boldsymbol{m}_\ell \odot \boldsymbol{q}_\ell \mid \text{data}]$. In the row labeled "$\mathbb{E}[\lambda_\ell \mid \text{data}]$", we report the posterior mean of the identity market price of risk, defined as $\mathbb{E}[\sum_j q_{\ell,j} m_{\ell,j} \mid \text{data}]$.

RP-PCs[12] and compare them with canonical PCs in Table 5.

According to Table 5, our method selects six identities when studying monthly data. PC1 and RP-PC1, despite explaining about 80% of the time-series return variation, are assigned to the same level identity and hence do not explain the cross-sectional spreads of mean returns. We also show that PC2 and PC4 have almost identical explanatory power as RP-PC3 and RP-PC4, respectively. In contrast, RP-PC2, RP-PC5, and PC5 are identified as uniquely important factors, although, as we show in Table A4 of the Appendix, the inclusion of PC5 is not robust to alternative prior specifications.

Investment-based asset pricing models predict that firm characteristics such as book-to-market, investment, and profitability are jointly determined by the same underlying state variables and hence correlated with expected returns (Berk et al., 1999; Cochrane, 1991; Zhang, 2005; Lin and Zhang, 2013). In particular, Berk et al. (1999) show that a firm's book-to-market ratio reflects the composition of its assets—the mix of assets in place and growth options—while investment represents the exercise of those options. A key implication is that

[12]The RP-PCA estimator is obtained by applying PCA to the matrix $\frac{1}{T}\boldsymbol{R}^\top\boldsymbol{R} + \gamma \cdot \widehat{\boldsymbol{\mu}}_r \widehat{\boldsymbol{\mu}}_r^\top$. When $\gamma = -1$, the solution is identical to that of standard PCA. We employ $\gamma = 10$ in estimating RP-PCs. In our empirical exercise, we further rotate the RP-PCs so that they are orthogonal in-sample, as is standard in PCA.

Table 5: PCs vs. RP-PCs

	Factor prob. within identities, $\boldsymbol{q}_\ell$						$p(\boldsymbol{m}_\ell \mid \text{data})$	$\mathbb{E}[\boldsymbol{\lambda} \mid \text{data}]$
	$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 4$	$\ell = 5$	$\ell = 6$	mean & 90% CI	
$\mathbf{1}_n$	0.017	0	0	0	0	0	0.123 (0.058, 0.193)	0.002
PC1	0.476	0	0	0	0	0	0.138 (0.065, 0.215)	0.066
PC2	0	0	0	0	0.020	0	-0.053 (-0.117, 0.011)	-0.001
PC3	0	0	0	0	0	0		0
PC4	0	0	0	0.007	0	0		0.001
PC5	0	0	0	0	0	1.000	-0.175 (-0.270, -0.082)	-0.175
RP-PC1	0.507	0	0	0	0	0	0.138 (0.065, 0.215)	0.070
RP-PC2	0	1.000	0	0	0	0	0.273 (0.200, 0.348)	0.273
RP-PC3	0	0	0	0	0.980	0	0.064 (-0.011, 0.139)	0.063
RP-PC4	0	0	0	0.993	0	0	0.117 (0.040, 0.191)	0.116
RP-PC5	0	0	1.000	0	0	0	0.376 (0.278, 0.473)	0.376
Ident. prob.	0.998	1.000	1.000	0.926	0.726	0.957		
$\mathbb{E}[\lambda_\ell \mid \text{data}]$	0.138	0.273	0.376	0.116	0.064	-0.175		

This table presents the Bayesian estimates of the first five PCs and RP-PCs (Lettau and Pelger, 2020), controlling for a common intercept ($\mathbf{1}_n$). The data are at the monthly frequency, spanning from January 1974 to December 2019. In empirical applications, all variables are standardized to have unit variance per period. The data description is in Section 3.1. We use Algorithm 1, with $L = 10$ and $\sigma_0 = 0.4/\sqrt{12}$, to estimate the posterior factor probabilities within identities ($\{\boldsymbol{q}_\ell\}_{\ell=1}^L$). The second last row reports the posterior identity probabilities ($\{\theta_\ell\}_{\ell=1}^L$). We display the identities with posterior probabilities above 50%. Conditional on $\{\boldsymbol{q}_\ell\}_{\ell=1}^L$, we estimate the posterior distribution of factor risk prices using the approach described in Section 1.3.2. We present both conditional and unconditional factor risk prices. In the column "$p(\boldsymbol{m}_\ell \mid \text{data})$", we report the factor risk prices conditional on the factors being selected (both the conditional posterior mean and 90% credible intervals). In the column "$\mathbb{E}[\boldsymbol{\lambda} \mid \text{data}]$", we display the unconditional posterior mean of the factor risk prices, defined as $\mathbb{E}[\sum_{\ell=1}^L \boldsymbol{m}_\ell \odot \boldsymbol{q}_\ell \mid \text{data}]$. In the row labeled "$\mathbb{E}[\lambda_\ell \mid \text{data}]$", we report the posterior mean of the identity market price of risk, defined as $\mathbb{E}[\sum_j q_{\ell,j} m_{\ell,j} \mid \text{data}]$.

value and investment factors should convey the same cross-sectional pricing information, even if their time-series returns differ. Different proxies for these underlying sources of risk have been proposed, including the Fama-French five-factor and q-factor models. Both models contain size, investment, and profitability factors. According to Panel A of Table 6, the two size (SMB and ME) and investment (CMA and IA) factors are highly correlated, with correlation coefficients above 0.9. In contrast, the two profitability factors, RMW and ROE, have a relatively low correlation of 0.67. But do these two sets of factors deliver heterogeneous cross-sectional pricing ability?

Table 7 suggests that the two size factors, SMB and ME, are not statistically distinguishable: they enter the same identity, with within-identity posterior probabilities of 43% and 57%, respectively.

Not only do CMA and IA enter the same identity, but the value factor, HML, is also

Table 6: Correlations among Fama-French Factors and Q-Factors

	MKT	SMB	HML	RMW	CMA	ME	IA	ROE
	Panel A. Time-series correlations among factors							
MKT	1.00							
SMB	0.22	1.00						
HML	-0.26	-0.03	1.00					
RMW	-0.26	-0.40	0.16	1.00				
CMA	-0.40	-0.03	0.68	0.06	1.00			
ME	0.22	0.97	0.01	-0.40	-0.01	1.00		
IA	-0.36	-0.11	0.67	0.16	0.91	-0.08	1.00	
ROE	-0.20	-0.39	-0.10	0.67	-0.05	-0.32	0.08	1.00
	Panel B. Cross-sectional correlations among $\text{Cov}(\boldsymbol{r}_t, \boldsymbol{f}_t)$							
MKT	1.00							
SMB	-0.45	1.00						
HML	-0.64	0.14	1.00					
RMW	0.14	-0.73	0.25	1.00				
CMA	-0.69	0.24	0.95	0.14	1.00			
ME	-0.46	1.00	0.18	-0.69	0.28	1.00		
IA	-0.61	0.10	0.96	0.29	0.98	0.14	1.00	
ROE	0.43	-0.72	-0.22	0.81	-0.25	-0.67	-0.12	1.00

This table presents the correlations among the Fama-French five factors (Fama and French, 2015) and q-factors (Hou et al., 2015). The data are at the monthly frequency, spanning from January 1974 to December 2019. In empirical applications, all variables are standardized to have unit variance per period. The data description is in Section 3.1. In Panel A, we report the time-series correlations among factor returns, whereas in Panel B, we show the cross-sectional correlations among their covariance loadings $\text{Cov}(\boldsymbol{r}_t, \boldsymbol{f}_t)$.

included in the same identity as the investment factors, consistent with the theoretical prediction of Berk et al. (1999) that both characteristics reflect the same underlying growth-option mix. Although the time-series correlation between HML and the two investment factors is moderate at about 0.67, the cross-sectional correlation in their covariance exposures, as shown in Panel B of Table 6, is close to 100%. Thus, HML has an almost identical cross-sectional pricing ability to that of the CMA and IA factors.

Nevertheless, the profitability factors are not alike. In the second identity, only ROE is selected as a strong pricing factor. We repeat the same analysis in Table A5 for two alternative priors and confirm that the major difference between these two factor models lies only in the profitability factor.

The shared identity of value and investment factors has a natural theoretical foundation. In the model of Berk et al. (1999), a firm's market value reflects the sum of assets in place and growth options, and investment is the exercise of those options. The book-to-market ratio serves as a sufficient statistic for the firm's mix of assets in place relative to growth

Table 7: Fama-French Factors vs. Q-Factors

	Factor prob. within identities, $\boldsymbol{q}_\ell$				$p(\boldsymbol{m}_\ell \mid \text{data})$	$\mathbb{E}[\boldsymbol{\lambda} \mid \text{data}]$
	$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 4$	mean & 90% CI	
$\mathbf{1}_n$	0.977	0	0.001	0	0.189 (0.120, 0.259)	0.185
MKT	0.023	0	0	0	0.213 (0.135, 0.291)	0.005
SMB	0	0	0.431	0	0.090 (0.004, 0.175)	0.039
HML	0	0	0	0.958	0.139 (0.059, 0.219)	0.133
RMW	0	0	0.001	0		0
CMA	0	0	0	0.015	0.101 (0.044, 0.159)	0.002
ME	0	0	0.566	0	0.091 (0.004, 0.179)	0.052
IA	0	0	0	0.026	0.109 (0.047, 0.173)	0.003
ROE	0	1.000	0	0	0.259 (0.169, 0.347)	0.259
Ident. prob.	1.000	1.000	0.887	0.983		
$\mathbb{E}[\lambda_\ell \mid \text{data}]$	0.190	0.259	0.091	0.138		

This table presents the Bayesian estimates of the Fama-French five factors (Fama and French, 2015) and q-factors (Hou et al., 2015), controlling for a common intercept ($\mathbf{1}_n$). The data are at the monthly frequency, spanning from January 1974 to December 2019. In empirical applications, all variables are standardized to have unit variance per period. The data description is in Section 3.1. We use Algorithm 1, with $L = 10$ and $\sigma_0 = 0.4/\sqrt{12}$, to estimate the posterior factor probabilities within identities ($\{\boldsymbol{q}_\ell\}_{\ell=1}^L$). The second last row reports the posterior identity probabilities ($\{\theta_\ell\}_{\ell=1}^L$). We display the identities with posterior probabilities above 50%. Conditional on $\{\boldsymbol{q}_\ell\}_{\ell=1}^L$, we estimate the posterior distribution of factor risk prices using the approach described in Section 1.3.2. We present both conditional and unconditional factor risk prices. In the column "$p(\boldsymbol{m}_\ell \mid \text{data})$", we report the factor risk prices conditional on the factors being selected (both the conditional posterior mean and 90% credible intervals). In the column "$\mathbb{E}[\boldsymbol{\lambda} \mid \text{data}]$", we display the unconditional posterior mean of the factor risk prices, defined as $\mathbb{E}[\sum_{\ell=1}^L \boldsymbol{m}_\ell \odot \boldsymbol{q}_\ell \mid \text{data}]$. In the row labeled "$\mathbb{E}[\lambda_\ell \mid \text{data}]$", we report the posterior mean of the identity market price of risk, defined as $\mathbb{E}[\sum_j q_{\ell,j} m_{\ell,j} \mid \text{data}]$.

options: high book-to-market firms have exhausted most of their growth options and invest conservatively, while low book-to-market firms are option-rich and invest aggressively. Both characteristics therefore sort firms along the same underlying dimension of risk, so that long-short portfolios formed on either dimension generate near-identical cross-sectional risk exposures. The investment Euler equation formalizes a complementary channel: the firm's discount rate simultaneously determines both Tobin's q (closely related to book-to-market) and the optimal investment rate (Cochrane, 1991; Zhang, 2005; Lin and Zhang, 2013; Hou et al., 2015). The finding that HML, CMA, and IA form a single pricing identity provides direct cross-sectional evidence for these theoretical predictions.

3.3 Three Types of Factor Models

We now turn to more extensive comparisons among three categories of factors: macroeconomic, statistical, and characteristic-based. In the following analysis, we include all factors in each category and examine their identities.

Table 8 presents the comparisons between statistical and characteristic-based factors at the monthly frequency. In Panel A, we use the Fama-French five factors (Fama and French, 2015), whereas their hedged returns (Daniel et al., 2020) are employed in Panel B. In both panels, none of the characteristic-based factors are selected in any identity. This observation holds even when we consider other values of the prior parameter σ_0, as we show in Table A6 in the Appendix. Several latent factors are robust explanators for mean excess returns, including RP-PC2, RP-PC5, and (PC4, RP-PC4). The only observable factor with a high posterior probability is the market factor, or its denoised version (MKT$\star$). More specifically, the market portfolio is a level factor, joining the same identity as PC1 and RP-PC1, whereas its denoised version (MKT$\star$), as Panel B indicates, displaces the other level factors and forms a unique identity in the SDF.

Tables A12–A14 augment the PCs with characteristic-based factors that belong to the same "style." In Table A12, we jointly examine the PCs, RP-PCs, ten value factors, and two market portfolios. As expected, all value factors consistently fall into the same pricing group—best represented by those sorted on the book-to-market ratio, the cash-flow-to-market ratio, and cash-flow duration. Notably, the second conventional PC yields pricing implications indistinguishable from this value group. Table A13 conducts the analogous exercise for nine factors constructed using technical signals. When σ_0 is large, two reversal signals enter the same group as the second canonical PC and the third RP-PC. Moreover, the industry-level momentum-and-reversal factor appears to deliver pricing implications that are unique relative to those of the PCs. Table A14 covers the remaining styles: once the PCs are included, none of the characteristic-based factors from these other styles provide incremental pricing information.

To address the concern that the risk-premia PCs may mechanically crowd out observable factors (because they are constructed specifically to explain mean excess returns) we re-estimate the model using only conventional PCs alongside the characteristic-based factors. Table A7 shows that the conclusions are largely unchanged. In Panel A, where we use the original observable factors of Panel A of Table 8, the selected identities are still overwhelmingly statistical: the first identity is mainly a level component, while the remaining identities are captured by PC3, PC4, and PC5. The characteristic-based factors receive virtually no

Table 8: Statistical vs. Characteristic-Based Factors

Panel A

	Factor prob. within identities, $\boldsymbol{q}_\ell$				$p(\boldsymbol{m}_\ell \mid \text{data})$	$\mathbb{E}[\boldsymbol{\lambda} \mid \text{data}]$
	$\ell=1$	$\ell=2$	$\ell=3$	$\ell=4$	mean & 90% CI	
$\mathbf{1}_n$	0.007	0	0	0		0.001
PC1	0.181	0	0	0	0.139 (0.065, 0.215)	0.025
PC2	0	0	0	0		-0.005
PC3	0	0	0	0		0.003
PC4	0	0	0	0.829	0.101 (0.028, 0.173)	0.086
PC5	0	0	0	0		-0.046
RP-PC1	0.200	0	0	0	0.139 (0.065, 0.215)	0.028
RP-PC2	0	1.000	0	0	0.223 (0.149, 0.300)	0.224
RP-PC3	0	0	0	0		0.010
RP-PC4	0	0	0	0.170	0.095 (0.025, 0.164)	0.018
RP-PC5	0	0	1.000	0	0.262 (0.190, 0.332)	0.267
MKT	0.613	0	0	0	0.140 (0.065, 0.217)	0.086
SMB	0	0	0	0		-0.001
HML	0	0	0	0		0
RMW	0	0	0	0		0
CMA	0	0	0	0		0
MOM	0	0	0	0		0.005
ME	0	0	0	0		-0.001
IA	0	0	0	0		0
ROE	0	0	0	0		0.001
EG	0	0	0	0		0
MGMT	0	0	0	0		0
PERF	0	0	0	0		0.003
PEAD	0	0	0	0		0.007
FIN	0	0	0	0		0
Ident. prob.	0.998	1.000	1.000	0.832		
$\mathbb{E}[\lambda_\ell \mid \text{data}]$	0.139	0.223	0.262	0.100		

Panel B

	Factor prob. within identities, $\boldsymbol{q}_\ell$				$p(\boldsymbol{m}_\ell \mid \text{data})$	$\mathbb{E}[\boldsymbol{\lambda} \mid \text{data}]$
	$\ell=1$	$\ell=2$	$\ell=3$	$\ell=4$	mean & 90% CI	
$\mathbf{1}_n$	0	0	0	0		0
PC1	0	0	0	0		0
PC2	0	0	0	0		0.003
PC3	0	0	0	0		-0.002
PC4	0	0.001	0	0		0.001
PC5	0	0	0	0		-0.012
RP-PC1	0	0	0	0		0
RP-PC2	0	0	0	1.000	0.129 (0.030, 0.224)	0.130
RP-PC3	0	0	0	0		-0.004
RP-PC4	0	0.999	0	0	0.166 (0.082, 0.246)	0.167
RP-PC5	0	0	1.000	0	0.272 (0.197, 0.344)	0.280
MKT⋆	1.000	0	0	0	0.284 (0.131, 0.434)	0.284
SMB⋆	0	0	0	0		0.003
HML⋆	0	0	0	0		0.002
RMW⋆	0	0	0	0		-0.016
CMA⋆	0	0	0	0		0.001
MOM	0	0	0	0		-0.002
ME	0	0	0	0		0.001
IA	0	0	0	0		0
ROE	0	0	0	0		-0.002
EG	0	0	0	0		0
MGMT	0	0	0	0		0
PERF	0	0	0	0		-0.006
PEAD	0	0	0	0		-0.005
FIN	0	0	0	0		0
Ident. prob.	0.998	0.980	0.999	0.913		
$\mathbb{E}[\lambda_\ell \mid \text{data}]$	0.284	0.166	0.272	0.129		

This table presents the Bayesian estimates of the statistical and characteristic-based factors. The data are at the monthly frequency, spanning from January 1974 to December 2019. In empirical applications, all variables are standardized to have unit variance per period. The data description is in Section 3.1. In Panel A, we use the Fama-French five factors (Fama and French, 2015), whereas the hedged returns (Daniel et al., 2020) of the Fama-French five factors are employed in Panel B. We use Algorithm 1, with $L = 10$ and $\sigma_0 = 0.4/\sqrt{12}$, to estimate the posterior factor probabilities within identities ($\{\boldsymbol{q}_\ell\}_{\ell=1}^L$). The second last row reports the posterior identity probabilities ($\{\theta_\ell\}_{\ell=1}^L$). We display the identities with posterior probabilities above 50%. Conditional on $\{\boldsymbol{q}_\ell\}_{\ell=1}^L$, we estimate the posterior distribution of factor risk prices using the approach described in Section 1.3.2. We present both conditional and unconditional factor risk prices. In the column "$p(\boldsymbol{m}_\ell \mid \text{data})$", we report the factor risk prices conditional on the factors being selected (both the conditional posterior mean and 90% credible intervals). In the column "$\mathbb{E}[\boldsymbol{\lambda} \mid \text{data}]$", we display the unconditional posterior mean of the factor risk prices, defined as $\mathbb{E}[\sum_{\ell=1}^L \boldsymbol{m}_\ell \odot \boldsymbol{q}_\ell \mid \text{data}]$. In the row labeled "$\mathbb{E}[\lambda_\ell \mid \text{data}]$", we report the posterior mean of the identity market price of risk, defined as $\mathbb{E}[\sum_j q_{\ell,j} m_{\ell,j} \mid \text{data}]$.

posterior support in this horse race. In Panel B, where we use the hedged factors (as in Panel B of Table 8), the denoised market factor MKT⋆ is selected with probability one as its own identity (displacing both the first PC and the intercept), while the other identities are again primarily represented by standard PCs. The only nonmarket observable factor that receives any noticeable support is HML⋆, with a posterior probability of about 2%, but it is selected within the same identity as PC2 (which has a posterior probability exceeding 94%) rather than as an independent source of pricing information.

Given that statistical factors evidently drive out most characteristic-based factors, do macroeconomic factors contain any incremental information beyond PCs? It depends on how we measure macroeconomic risk.

In Table A9 of the Appendix, we conduct a horse-racing exercise between statistical factors and 11 nontraded macro factors whose growth rates are measured at a quarterly frequency. We estimate the five largest PCs (RP-PCs) of quarterly asset returns.[13] Across all three specifications, only the HKM intermediary factor is selected into the first identity, joining the same identity as the common intercept $\mathbf{1}_n$, PC1, and RP-PC1. That is, the HKM factor is a level factor, consistent with the findings in Table 4. All other macroeconomic fundamentals, including several measures of production, consumption, and unemployment, as well as TFP growth and aggregate liquidity, are not selected in any identity.

Although the link between asset returns and the macroeconomy appears weak at the quarterly frequency, past research often shows that returns' exposures to macro fundamentals, when measured over multiple periods, tend to explain the cross-section of mean returns significantly better than in short horizons.[14] Motivated by this observation, we construct the 12-quarter cumulative growth rates of macro variables, following Parker and Julliard (2005), and compare them with the statistical factors in terms of their ability to explain quarterly average returns.

Table 9 shows that six identities are selected, with the first five consisting of statistical factors and, in some specifications, the HKM factor as a level factor. Unlike the macro fundamentals measured at the quarterly frequency studied previously, we now recover a set of significant macro factors (see column $\ell = 6$): the 12-quarter cumulative growth rates of GDP, industrial production, and nondurable plus service consumption constitute a unique identity, with posterior factor probabilities of 12%, 4%, and 84%, respectively. Their risk prices are also significantly positive and of similar magnitudes. For example, the risk price of nondurable plus service consumption growth has a conditional posterior mean of 0.28, with a 90% credible interval of (0.10, 0.48). Therefore, our estimates imply that when investors' marginal utilities are high today, they expect the cumulative growth of several key economic fundamentals to be low over the next three years. The asset pricing model is not sparse in the space of statistical and macro factors: 11 out of 22 candidate factors are selected, whereas sparsity arises only at the identity level.

[13] Note that the quarterly PCs are different from those in previous tables based on monthly return data.

[14] See, for example, Parker and Julliard (2005), Jagannathan and Wang (2007), Ortu et al. (2013), Bandi and Tamoni (2023), and Bryzgalova et al. (2025a).

We conduct robustness checks in Table A8 under two alternative prior formulations. In both specifications, we consistently identify a common set of macro fundamentals, with the other five identities composed of statistical and HKM factors. The only distinction is that when the prior identity-level Sharpe ratio is 0.5, the 12-quarter growth rate of aggregate investment emerges as significant, replacing consumption growth in the macro identity. This observation echoes the simulated example in Table 3. When we vary the prior specification, different factors within the same identity may appear significant, because our approach selects some, but not all, of the useful factors in a given identity.

What is the mechanism behind the seemingly interchangeable pricing ability of several macro factors? Angeletos et al. (2020) use a novel identification scheme in a vector autoregressive (VAR) model to show that five macroeconomic quantities (unemployment, output, hours worked, consumption, and investment) are mainly driven by the same shock at the business-cycle frequencies. Bryzgalova et al. (2025b), leveraging the information in a large cross-section of asset returns, further suggest that an almost identical asset return shock drives the dynamics of these economic fundamentals over two- to three-year horizons. The findings from the previous two papers imply that the covariance exposures of these variables are proportional to one another. In the data, we confirm that the cross-sectional correlations among the covariance exposures of GDP, industrial production, nondurable plus service consumption, and investment (i.e., $\text{Cov}(\boldsymbol{r}_{t-1,t}, \boldsymbol{f}_{t-1,t+12})$) range from 89% to 94%, which explains the near-homogeneous pricing behaviour of these four macro factors.

Can the characteristic-based factors subsume the pricing information contained in the macro factors? To answer this question, we compare macro fundamentals with characteristic-based factors. Table 10 shows that the same set of macro factors (namely, the cumulative growth rates of GDP, industrial production, and consumption) remains jointly selected by our Bayesian method. The selection of macro quantities is robust to alternative prior specifications, as reported in Table A10 of the Appendix. Although TFP growth commands a non-trivial risk price, with an unconditional posterior mean of 0.15, it is not selected in any pricing identity. Moreover, the covariance exposures to TFP growth, even when measured at the three-year horizon, are small: the L_2 norm of $\text{Cov}(\boldsymbol{r}_{t-1,t}, \text{TFP}_{t-1,t+12})$ is only one-tenth of those of other selected macro quantities. Therefore, TFP growth is likely a weak factor

Table 9: Statistical vs. Macro Factors

	Factor probs. within identities, $\boldsymbol{q}_\ell$						$p(\boldsymbol{m}_\ell \mid \text{data})$	$\mathbb{E}[\boldsymbol{\lambda} \mid \text{data}]$
	$\ell=1$	$\ell=2$	$\ell=3$	$\ell=4$	$\ell=5$	$\ell=6$	mean & 90% CI	
$\mathbf{1}_n$	0.891	0	0	0	0	0	0.177 (0.042, 0.320)	0.158
PC1	0.047	0	0	0	0	0	0.196 (0.045, 0.353)	0.009
PC2	0	0	0	0	0	0		0
PC3	0	0	0	0	0	0		0.001
PC4	0	0	0	0	0.075	0	0.145 (0.001, 0.291)	0.011
PC5	0	0	0	0	0	0		-0.016
RP-PC1	0.057	0	0	0	0	0	0.196 (0.045, 0.353)	0.011
RP-PC2	0	1.000	0	0	0	0	0.345 (0.189, 0.504)	0.346
RP-PC3	0	0	0	1.000	0	0	0.177 (0.021, 0.328)	0.177
RP-PC4	0	0	0	0	0.925	0	0.159 (0.006, 0.309)	0.148
RP-PC5	0	0	1.000	0	0	0	0.252 (0.097, 0.409)	0.254
Nondurable (PJ12)	0	0	0	0	0	0		0.001
GDP (PJ12)	0	0	0	0	0	0.119	0.215 (0.056, 0.399)	0.030
Industrial production (PJ12)	0	0	0	0	0	0.039	0.209 (0.033, 0.432)	0.012
Nondurable+service (PJ12)	0	0	0	0	0	0.837	0.282 (0.095, 0.472)	0.245
Durable (PJ12)	0	0	0	0	0	0		0.001
Investment (PJ12)	0	0	0	0	0	0.002		0.002
Hours worked (PJ12)	0	0	0	0	0	0.002		0.002
Unemp. rate (PJ12)	0	0	0	0	0	0		-0.001
TFP (PJ12)	0	0	0	0	0	0		0.040
Liquidity-ntr	0	0	0	0	0	0		0
HKM-ntr	0.004	0	0	0	0	0		0.001
Ident. prob.	0.990	0.999	0.915	0.868	0.757	0.886		
$\mathbb{E}[\lambda_\ell \mid \text{data}]$	0.180	0.345	0.252	0.177	0.158	0.271		

This table presents the Bayesian estimates of the statistical and macro factors. The data are at the quarterly frequency, spanning from 1974Q1 to 2019Q4. In empirical applications, all variables are standardized to have unit variance per period. The data description is in Section 3.1. Macro factors with the notation "(PJ12)" are defined as the 12-quarter cumulative growth rates of the underlying economic variables between quarter $t-1$ and quarter $t+12$, following Parker and Julliard (2005). We use Algorithm 1, with $L=10$ and $\sigma_0 = 0.4/\sqrt{4}$, to estimate the posterior factor probabilities within identities ($\{\boldsymbol{q}_\ell\}_{\ell=1}^L$). The second last row reports the posterior identity probabilities ($\{\theta_\ell\}_{\ell=1}^L$). We display the identities with posterior probabilities above 50%. Conditional on $\{\boldsymbol{q}_\ell\}_{\ell=1}^L$, we estimate the posterior distribution of factor risk prices using the approach described in Section 1.3.2. We present both conditional and unconditional factor risk prices. In the column "$p(\boldsymbol{m}_\ell \mid \text{data})$", we report the factor risk prices conditional on the factors being selected (both the conditional posterior mean and 90% credible intervals). In the column "$\mathbb{E}[\boldsymbol{\lambda} \mid \text{data}]$", we display the unconditional posterior mean of the factor risk prices, defined as $\mathbb{E}[\sum_{\ell=1}^L \boldsymbol{m}_\ell \odot \boldsymbol{q}_\ell \mid \text{data}]$. In the row labeled "$\mathbb{E}[\lambda_\ell \mid \text{data}]$", we report the posterior mean of the identity market price of risk, defined as $\mathbb{E}[\sum_j q_{\ell,j} m_{\ell,j} \mid \text{data}]$.

that is almost uncorrelated with equity returns.

In addition to the macro factors, the momentum and performance (PERF) mispricing factors jointly enter the same pricing identity, with within-identity posterior probabilities of 24% and 76%, respectively. That is, conditional on other control variables, we cannot distinguish the pricing information embedded in the momentum and performance mispricing factors. The denoised market and size factors (MKT$\star$ and SMB$\star$), following Daniel et al. (2020), form two additional factor identities that displace the traditional market factor and

Table 10: Characteristic-Based vs. Macro Factors

	Factor probs. within identities, $\boldsymbol{q}_\ell$				$p(\boldsymbol{m}_\ell \mid \text{data})$	$\mathbb{E}[\boldsymbol{\lambda} \mid \text{data}]$
	$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 4$	mean & 90% CI	
$\mathbf{1}_n$	0	0	0	0		0
MKT	0	0	0	0		0
SMB	0	0	0	0		0
HML	0	0	0	0		0.001
RMW	0	0	0	0		0
CMA	0	0	0	0		0.001
MOM	0	0	0	0.244	0.216 (0.048, 0.397)	0.054
MKT$\star$	1.000	0	0	0	0.390 (0.149, 0.628)	0.390
SMB$\star$	0	1.000	0	0	0.340 (0.139, 0.540)	0.343
HML$\star$	0	0	0	0		0.002
RMW$\star$	0	0	0	0		-0.002
CMA$\star$	0	0	0	0		0.011
ME	0	0	0	0		0.001
IA	0	0	0	0		0.001
ROE	0	0	0	0		0
EG	0	0	0	0		0
MGMT	0	0	0	0		0
PERF	0	0	0	0.756	0.216 (0.044, 0.391)	0.165
PEAD	0	0	0	0		0.007
FIN	0	0	0	0		0
Nondurable (PJ12)	0	0	0	0		0.001
GDP (PJ12)	0	0	0.052	0	0.261 (0.060, 0.499)	0.020
Industrial production (PJ12)	0	0	0.012	0	0.255 (0.035, 0.554)	0.010
Nondurable+service (PJ12)	0	0	0.935	0	0.350 (0.099, 0.596)	0.345
Durable (PJ12)	0	0	0	0		0.005
Investment (PJ12)	0	0	0	0		0.002
Hours worked (PJ12)	0	0	0	0		0.001
Unemp. rate (PJ12)	0	0	0	0		0
TFP (PJ12)	0	0	0	0		0.153
Liquidity-ntr	0	0	0	0		0
HKM-ntr	0	0	0	0		0
Ident. prob.	0.998	0.968	0.868	0.946		
$\mathbb{E}[\lambda_\ell \mid \text{data}]$	0.390	0.340	0.344	0.216		

This table presents the Bayesian estimates of the characteristic-based and macro factors. The data are at the quarterly frequency, spanning from 1974Q1 to 2019Q4. In empirical applications, all variables are standardized to have unit variance per period. The data description is in Section 3.1. Macro factors with the notation "(PJ12)" are defined as the 12-quarter cumulative growth rates of the underlying economic variables between quarter $t-1$ and quarter $t+12$, following Parker and Julliard (2005). We use Algorithm 1, with $L = 10$ and $\sigma_0 = 0.4/\sqrt{4}$, to estimate the posterior factor probabilities within identities ($\{\boldsymbol{q}_\ell\}_{\ell=1}^L$). The second last row reports the posterior identity probabilities ($\{\theta_\ell\}_{\ell=1}^L$). We display the identities with posterior probabilities above 50%. Conditional on $\{\boldsymbol{q}_\ell\}_{\ell=1}^L$, we estimate the posterior distribution of factor risk prices using the approach described in Section 1.3.2. We present both conditional and unconditional factor risk prices. In the column "$p(\boldsymbol{m}_\ell \mid \text{data})$", we report the factor risk prices conditional on the factors being selected (both the conditional posterior mean and 90% credible intervals). In the column "$\mathbb{E}[\boldsymbol{\lambda} \mid \text{data}]$", we display the unconditional posterior mean of the factor risk prices, defined as $\mathbb{E}[\sum_{\ell=1}^L \boldsymbol{m}_\ell \odot \boldsymbol{q}_\ell \mid \text{data}]$. In the row labeled "$\mathbb{E}[\lambda_\ell \mid \text{data}]$", we report the posterior mean of the identity market price of risk, defined as $\mathbb{E}[\sum_j q_{\ell,j} m_{\ell,j} \mid \text{data}]$.

the two size factors proposed by Fama and French (2015) and Hou et al. (2015). Thus, purifying the characteristic-based factors by removing unpriced components not only improves their investment performance, in terms of mean-variance efficiency, but also enhances their ability to explain the cross-section of mean returns relative to the original sorted factors.

Finally, we conduct a comprehensive comparison of statistical, characteristic-based, and macroeconomic factors. Table 11 indicates that five identities are selected, with the first four composed of the denoised market factor (MKT$\star$) and (RP-)PCs. Most strikingly, even after controlling for a large set of statistical and characteristic-based factors, a unique identity of macro fundamentals—growth rates of GDP, industrial production, and consumption measured over a three-year horizon—is recovered. By contrast, we still find no strong evidence supporting the selection of characteristic-based factors. These findings are not sensitive to the choice of prior, as suggested by Table A11.

Next, using the same model specification given in Table 11, we estimate the posterior density of the annualized Sharpe ratio implied by the linear SDF model and present the results in Figure 2 for three values of the prior parameter. Even though 40 factors are considered ex ante, the ex post model-implied Sharpe ratio is not excessive: when the prior identity-level Sharpe ratio is 0.4, the unconditional posterior mean is 1.49 per annum, with a 90% credible interval of $(1.05, 1.96)$. Moreover, the posterior distributions are stable across the three prior specifications.

We further decompose the implied Sharpe ratio to better understand which types of economic risks drive the variation in the latent pricing kernel. Specifically, we decompose the SDF into tradable and nontradable components: $M_t = M_t^{\text{tradable}} + M_t^{\text{macro}}$, where M_t^{tradable} and M_t^{macro} denote the SDF components constructed from 29 tradable (ten statistical and 19 characteristic-based) and 11 macroeconomic factors, respectively. As we show in Table 12, the tradable component accounts for 52%–69% of the SDF, with annualized Sharpe ratios of 1.01–1.19. Perhaps more surprisingly, the macroeconomic SDF component is also sizable, explaining 18%–32% of the SDF. There is also substantial common priced information across the two types of factors, as the two SDF components are positively correlated. As indicated by the column "Cov." of Table 12, their covariance terms explain the remaining 13%–16% of the SDF variation.

Table 11: Statistical vs. Characteristic-Based vs. Macro Factors

	Factor probs. within identities, $\boldsymbol{q}_\ell$					$p(\boldsymbol{m}_\ell \mid \text{data})$	$\mathbb{E}[\boldsymbol{\lambda} \mid \text{data}]$
	$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 4$	$\ell = 5$	mean & 90% CI	
$\mathbf{1}_n$	0	0	0	0	0		0
PC1	0	0	0	0	0		0
PC2	0	0	0	0	0		0
PC3	0	0	0	0	0		0
PC4	0	0	0.001	0	0		0.001
PC5	0	0	0	0	0		-0.005
RP-PC1	0	0	0	0	0		0
RP-PC2	0	1.000	0	0	0	0.253 (0.066, 0.437)	0.253
RP-PC3	0	0	0	0	0		0.001
RP-PC4	0	0	0.999	0	0	0.295 (0.113, 0.491)	0.296
RP-PC5	0	0	0	1.000	0	0.277 (0.115, 0.444)	0.279
MKT	0	0	0	0	0		0
SMB	0	0	0	0	0		0
HML	0	0	0	0	0		0
RMW	0	0	0	0	0		0
CMA	0	0	0	0	0		0
MOM	0	0	0	0	0		0
MKT$\star$	1.000	0	0	0	0	0.392 (0.132, 0.659)	0.392
SMB$\star$	0	0	0	0	0		-0.001
HML$\star$	0	0	0	0	0		0
RMW$\star$	0	0	0	0	0		-0.001
CMA$\star$	0	0	0	0	0		0
ME	0	0	0	0	0		0
IA	0	0	0	0	0		0
ROE	0	0	0	0	0		0
EG	0	0	0	0	0		0
MGMT	0	0	0	0	0		0
PERF	0	0	0	0	0		0
PEAD	0	0	0	0	0		0.001
FIN	0	0	0	0	0		0
Nondurable (PJ12)	0	0	0	0	0		0.001
GDP (PJ12)	0	0	0	0	0.072	0.201 (0.004, 0.422)	0.018
Industrial production (PJ12)	0	0	0	0	0.031	0.193 (-0.007, 0.464)	0.009
Nondurable+service (PJ12)	0	0	0	0	0.893	0.264 (0.007, 0.526)	0.242
Durable (PJ12)	0	0	0	0	0		0.001
Investment (PJ12)	0	0	0	0	0.002		0.002
Hours worked (PJ12)	0	0	0	0	0.002		0.001
Unemp. rate (PJ12)	0	0	0	0	0		-0.001
TFP (PJ12)	0	0	0	0	0		0.080
Liquidity-ntr	0	0	0	0	0		0
HKM-ntr	0	0	0	0	0		0
Ident. prob.	0.995	0.945	0.954	0.925	0.816		
$\mathbb{E}[\lambda_\ell \mid \text{data}]$	0.392	0.253	0.295	0.277	0.257		

This table presents the Bayesian estimates of the statistical, characteristic-based, and macro factors. The data are at the quarterly frequency, spanning from 1974Q1 to 2019Q4. In empirical applications, all variables are standardized to have unit variance per period. The data description is in Section 3.1. Macro factors with the notation "(PJ12)" are defined as the 12-quarter cumulative growth rates of the underlying economic variables between quarter $t-1$ and quarter $t+12$, following Parker and Julliard (2005). We use Algorithm 1, with $L = 10$ and $\sigma_0 = 0.4/\sqrt{4}$, to estimate the posterior factor probabilities within identities ($\{\boldsymbol{q}_\ell\}_{\ell=1}^L$). The second last row reports the posterior identity probabilities ($\{\theta_\ell\}_{\ell=1}^L$). We display the identities with posterior probabilities above 50%. Conditional on $\{\boldsymbol{q}_\ell\}_{\ell=1}^L$, we estimate the posterior distribution of factor risk prices using the approach described in Section 1.3.2. We present both conditional and unconditional factor risk prices. In the column "$p(\boldsymbol{m}_\ell \mid \text{data})$", we report the factor risk prices conditional on the factors being selected (both the conditional posterior mean and 90% credible intervals). In the column "$\mathbb{E}[\boldsymbol{\lambda} \mid \text{data}]$", we display the unconditional posterior mean of the factor risk prices, defined as $\mathbb{E}[\sum_{\ell=1}^L \boldsymbol{m}_\ell \odot \boldsymbol{q}_\ell \mid \text{data}]$. In the row labeled "$\mathbb{E}[\lambda_\ell \mid \text{data}]$", we report the posterior mean of the identity market price of risk, defined as $\mathbb{E}[\sum_j q_{\ell,j} m_{\ell,j} \mid \text{data}]$.

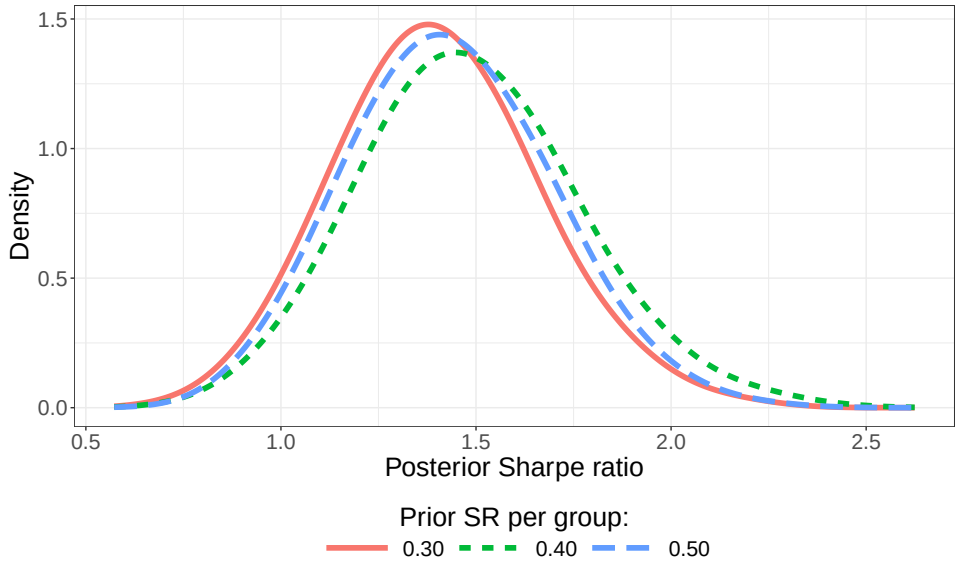

Figure 2: Posterior Distribution of the SDF-Implied Sharpe Ratio

This figure plots the posterior density of the annualized Sharpe ratio implied by the linear SDF model for various values of the annualized identity-level prior Sharpe ratio. We include ten statistical factors, 19 tradable characteristic-based factors, and 11 nontradable macroeconomic factors, as listed in Table 11. The data are at the quarterly frequency, spanning 1974Q1–2019Q4.

In summary, by systematically comparing the three types of factors, we show that statistical factors are consistently the most important for cross-sectional pricing. Controlling for statistical factors, we find no strong evidence supporting characteristic-based factors. Interestingly, a set of macro quantities related to aggregate output, industrial production, consumption, and, in some specifications, investment are robustly identified in the data when their covariance loadings are properly measured. Level factors—whose covariances with asset returns are approximately constant in the cross-section—are common in the data: for instance, the first (risk-premia) PC, the market portfolio, and the HKM intermediary factor all exhibit this level behavior. Under our estimation framework, their risk prices are statistically different from zero and therefore contribute to the SDF, although they are subject to identification problems under other standard estimation methods.

When interpreting these horse-racing results, we need to keep in mind that the observed exclusion of characteristic-based factors does not cast doubt on their usefulness. In fact, statistical factors, especially RP-PCs, are designed to capture the time-series and cross-sectional properties of asset returns. As more (RP-)PCs are included, observable factors are

Table 12: Sharpe Ratio Decomposition in the SDF: Tradable vs. Nontradable

	Panel A. $\sigma_0 = 0.3/\sqrt{4}$			Panel B. $\sigma_0 = 0.4/\sqrt{4}$			Panel C. $\sigma_0 = 0.5/\sqrt{4}$		
	$\text{SR}_f^{\text{tradable}}$	$\text{SR}_f^{\text{macro}}$	Cov.	$\text{SR}_f^{\text{tradable}}$	$\text{SR}_f^{\text{macro}}$	Cov.	$\text{SR}_f^{\text{tradable}}$	$\text{SR}_f^{\text{macro}}$	Cov.
Mean	1.150	0.576		1.190	0.647		1.008	0.795	
5%	0.795	0.203		0.817	0.213		0.706	0.289	
95%	1.528	0.976		1.592	1.130		1.330	1.321	
$\mathbb{E}\left[\frac{\text{SR}_f^2}{\text{SR}_m^2} \mid \text{data}\right]$	0.687	0.183	0.130	0.662	0.206	0.132	0.522	0.318	0.160

This table presents the variance decomposition of the SDF, $M_t = M_t^{\text{tradable}} + M_t^{\text{macro}}$. We include ten statistical factors, 19 tradable characteristic-based factors, and 11 nontradable macroeconomic factors, as listed in Table 11. M_t^{tradable} and M_t^{macro} denote the SDF components constructed from 29 tradable and 11 macroeconomic factors, respectively. $\text{SR}_f^{\text{tradable}}$ and $\text{SR}_f^{\text{macro}}$ are defined as $\sqrt{\text{Var}(M_t^{\text{tradable}})}$ and $\sqrt{\text{Var}(M_t^{\text{macro}})}$, respectively, and are all annualized. We report the posterior means and 90% credible intervals of $\text{SR}_f^{\text{tradable}}$ and $\text{SR}_f^{\text{macro}}$. The final row presents the fraction of the SDF variance explained by each of the three components, where the column "Cov." denotes the fraction explained by the covariance term, defined as $2 \times \text{Cov}(M_t^{\text{tradable}}, M_t^{\text{macro}})$. The data are at the quarterly frequency, spanning 1974Q1–2019Q4.

progressively driven out. By the same logic, we interpret the discovery of a macro-factor identity as indicating that the five largest latent factors do not fully subsume the pricing power of these macro fundamentals.

3.4 Factor Density and Identity Sparsity of the SDF

Our evidence so far consistently indicates that asset risk premia are driven by a small number of pricing identities. This finding is fully compatible with approaches that have instead documented very dense SDFs in the space of characteristics-based factors. To see why, note that, as shown by Dickerson et al. (2026), the Bayesian Model Averaging SDF (BMA-SDF) over the space of all possible (quadrillion) models of Bryzgalova et al. (2023) is identical to a linear combination of all factors with weights equal to their posterior risk prices, $\mathbb{E}[\boldsymbol{\lambda} \mid \text{data}]$. This construction yields a very dense SDF in the space of characteristic-based factors, echoing the findings of Kozak et al. (2020).

As shown in the last column of our Appendix Tables A15–A17, which report the mean and 90% credible intervals of $p(\boldsymbol{\lambda} \mid \text{data})$, our estimates also imply a dense SDF in the factor space, since most of the posterior means are different from zero. The leading identities still feature one or two proxies, while the forth or fifth identity requires almost all factors. Nevertheless, the same tables reveal that only a few identities carry substantial posterior probability. That is, the estimated SDF is dense in factor space but sparse in identity space. From a theoretical standpoint, this is not contradictory: as Dickerson et al. (2026) point out,

Table 13: Model Sparsity at the Identity Level

	Mean	0.5%	2.5%	5%	50%	95%	97.5%	99.5%
Panel A. Randomly draw 20 anomalies								
$\sigma_0 = 0.3/\sqrt{12}$	4.05	0.00	2.45	2.53	4.15	5.53	5.93	6.68
$\sigma_0 = 0.4/\sqrt{12}$	4.21	0.00	2.47	2.53	4.26	5.83	6.14	6.95
$\sigma_0 = 0.5/\sqrt{12}$	4.26	0.00	2.45	2.56	4.30	5.90	6.19	7.04
Panel B. Randomly draw 30 anomalies								
$\sigma_0 = 0.3/\sqrt{12}$	4.37	2.55	2.63	3.36	4.32	5.91	6.07	6.90
$\sigma_0 = 0.4/\sqrt{12}$	4.41	2.53	2.61	3.38	4.38	5.99	6.21	6.83
$\sigma_0 = 0.5/\sqrt{12}$	4.44	2.48	2.58	3.37	4.41	5.97	6.17	6.74

This table presents the distribution of model dimensionality at the group level across 1,000 random experiments. In each experiment, we randomly select 20 (Panel A) or 30 (Panel B) long-short anomalies, controlling for two market factors (MKT and MKT$\star$). For each experiment, we compute the posterior mean of the number of pricing groups selected by our Bayesian approach, $\sum_\ell \theta_\ell$. We repeat the experiment 1,000 times and report the mean and 0.5%, 2.5%, 5%, 50%, 95%, 97.5%, and 99.5% percentiles of $\sum_\ell \theta_\ell$. We use Algorithm 1, with $L = 10$ and $\sigma_0 \in \{0.3/\sqrt{12}, 0.4/\sqrt{12}, 0.5/\sqrt{12}\}$. The data are at the monthly frequency, spanning from January 1974 to December 2019. In empirical applications, all variables are standardized to have unit variance per period. The data description is in Section 3.1.

if multiple factors are noisy proxies for common underlying sources of risk, the BMA-SDF will optimally use them *all* (albeit with higher weights assigned to the more informative proxies) to maximize the signal-to-noise ratio and recover the true pricing implications of the latent SDF.

But a question remains: could our finding of an identity-sparse SDF be an artifact of the particular set of factors considered jointly in our estimation? Table 13 addresses this concern by conducting a random subsampling exercise. We randomly select 20 (Panel A) or 30 (Panel B) long-short anomalies, always controlling for two market factors (MKT and MKT$\star$). For each sampled set of factors, we estimate the number of identities present using our method. The table reports the distribution of the number of identities across these experiments, and the finding is stark: regardless of the prior, we nearly always find between three and six pricing identities (the centered 95% range spans from a minimum of 2.5 to a maximum of 6.2 identities). The factor zoo, it appears, is sparse in species despite being dense in animals.

4 Conclusion

The empirical dominance of dense, high-dimensional models for pricing the cross-section of expected returns has come at a cost: as the stochastic discount factor grows complex, the

economic narrative of priced risk disappears. This paper introduces the concept of a *factor identity* and develops the inferential tools to recover identities from data, reconciling the empirical reality of a dense SDF with the goal of economic interpretability.

Two factors share an identity when they are near-substitutes in terms of their pricing implications: their covariances with asset returns are sufficiently similar that the data cannot determine which one drives the cross-sectional variation in expected returns. This near-substitutability, together with the related challenges of weak and level factors, renders traditional inference unreliable. We develop a novel hierarchical Bayesian framework that solves all three identification challenges simultaneously, detects and groups near-substitute factors into pricing identities, ranks each member by its posterior probability of being the identity's single best proxy, and fully accounts for the estimation uncertainty in both expected returns and factor covariances that arises from the latent nature of the cross-sectional regression inputs.

The central empirical finding is that the SDF is dense in factors but sparse in identities. Applying our method to a comprehensive set of macroeconomic variables, characteristic-sorted portfolios, and statistical latent factors, only a handful of distinct economic identities emerge: a macroeconomic identity, best captured by long-horizon growth in output, consumption, and industrial production; a market identity, which pins down the overall level of risk premia; and a latent statistical identity, reflecting residual variation not yet attributable to observable fundamentals. The factor zoo is dense in animals, but its animals belong to just a handful of species.

This sparsity in identity space has practical implications that extend beyond description. A proposed state variable adds genuine economic content if and only if it introduces an identity not already spanned by the existing pricing groups. Our method thus provides the inferential apparatus against which new theories can be tested, restoring to the dense-SDF literature the capacity for economic dialogue that individual-factor selection has lost. We view this as an open direction: mapping emerging theoretical channels onto the identity structure documented here is a natural next step for empirical asset pricing research.

References

Adrian, T., E. Etula, and T. Muir (2014). Financial intermediaries and the cross-section of asset returns. *Fournal of Finance 69*(6), 2557–2596.

Anderson, E. W. and A.-R. Cheng (2016). Robust Bayesian portfolio choices. *Review of Financial Studies 29*(5), 1330–1375.

Angeletos, G.-M., F. Collard, and H. Dellas (2020). Business-cycle anatomy. *American Economic Review 110*(10), 3030–3070.

Avramov, D. (2002). Stock return predictability and model uncertainty. *Journal of Financial Economics 64*(3), 423–458.

Avramov, D., S. Cheng, L. Metzker, and S. Voigt (2023). Integrating factor models. *Journal of Finance, forthcoming*.

Bandi, F. M. and A. Tamoni (2023). Business-cycle consumption risk and asset prices. *Journal of Econometrics 237*(2), 105447.

Barillas, F. and J. Shanken (2017). Which alpha? *Review of Financial Studies 30*(4), 1316–1338.

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

Bartlett, M. S. (1957). A comment on D.V. Lindley's statistical paradox. *Biometrika 44*(3–4), 533–534.

Berk, J. B., R. C. Green, and V. Naik (1999). Optimal investment, growth options, and security returns. *Journal of Finance 54*(5), 1553–1607.

Blei, D. M., A. Kucukelbir, and J. D. McAuliffe (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association 112*(518), 859–877.

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., J. Huang, and C. Julliard (2025a). Consumption in asset returns. *Journal of Finance*. forthcoming, Available at SSRN 3783070.

Bryzgalova, S., J. Huang, and C. Julliard (2025b). Macro strikes back: Term structure of risk premia. *Available at SSRN 4752696*.

Chamberlain, G. and M. Rothschild (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica 51*, 1281–1304.

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

Cochrane, J. H. (1991). Production-based asset pricing and the link between stock returns and economic fluctuations. *Journal of Finance 46*(1), 209–237.

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

Cochrane, J. H. and J. Saa-Requejo (2000). Beyond arbitrage: Good-deal asset price bounds in incomplete markets. *Journal of Political Economy 108*(1), 79–119.

Connor, G. (1995). The three types of factor models: A comparison of their explanatory power. *Financial Analysts Journal 51*(3), 42–46.

Connor, G. and R. A. Korajczyk (1986). Performance measurement with the arbitrage pricing theory: A new framework for analysis. *Journal of Financial Economics 15*(3), 373–394.

Connor, G. and R. A. Korajczyk (1988). Risk and return in an equilibrium apt: Application of a new test methodology. *Journal of Financial Economics 21*(2), 255–289.

Da, R., S. Nagel, and D. Xiu (2025). The statistical limit of arbitrage. *Available at SSRN 4986807*.

Daniel, K., D. Hirshleifer, and L. Sun (2020). Short- and long-horizon behavioral factors. *Review of Financial Studies 33*(4), 1673–1736.

Daniel, K., L. Mota, S. Rottke, and T. Santos (2020). The cross-section of risk and returns. *Review of Financial Studies 33*(5), 1927–1979.

Dickerson, A., C. Julliard, and P. Mueller (2026). The co-pricing factor zoo. *Journal of Financial Economics*. forthcoming.

Fama, E. F. and K. R. French (1992, Jun). The cross-section of expected stock returns. *Journal of Finance 47*, 427–465.

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 J. D. MacBeth (1973). Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy 81*(3), 607–636.

Fan, J., Y. Fan, and J. Lv (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics 147*(1), 186–197.

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

Giglio, S., B. Kelly, and S. Kozak (2024). Equity term structures without dividend strips data. *Journal of Finance 79*(6), 4143–4196.

Giglio, S. and D. Xiu (2021). Asset pricing with omitted factors. *Journal of Political Economy 129*(7), 1947–1990.

Gospodinov, N., R. Kan, and C. Robotti (2014). Misspecification-robust inference in linear asset-pricing models with irrelevant risk factors. *The Review of Financial Studies 27*(7), 2139–2170.

Gospodinov, N., R. Kan, and C. Robotti (2019). Too good to be true? fallacies in evaluating risk factor models. *Journal of Financial Economics 132*(2), 451–471.

Haddad, V., S. Kozak, and S. Santosh (2020). Factor timing. *Review of Financial Studies 33*(5), 1980–2018.

Haddad, V. and T. Muir (2021). Do intermediaries matter for aggregate asset prices? *Journal of Finance 76*(6), 2719–2761.

Hansen, L. P. and R. Jagannathan (1991). Implications of security market data for models of dynamic economies. *Journal of Political Economy 99*(2), 225–262.

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

Harvey, C. R. and G. Zhou (1990). Bayesian inference in asset pricing tests. *Journal of Financial Economics 26*(2), 221–254.

He, A., D. Huang, J. Li, and G. Zhou (2023). Shrinking factor dimension: A reduced-rank approach. *Management Science 69*(9), 5501–5522.

He, Z., B. Kelly, and A. Manela (2017). Intermediary asset pricing: New evidence from many asset classes. *Journal of Financial Economics 126*(1), 1–35.

He, Z. and A. Krishnamurthy (2012). A model of capital and crises. *Review of Economic Studies 79*(2), 735–777.

He, Z. and A. Krishnamurthy (2013). Intermediary asset pricing. *American Economic Review 103*(2), 732–770.

Herskovic, B., A. Moreira, and T. Muir (2019). Hedging risk factors. Working paper, R&R at Review of Financial Studies.

Hou, K., H. Mo, C. Xue, and L. Zhang (2021). An augmented q-factor model with expected growth. *Review of Finance 25*(1), 1–41.

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

Huang, J. and R. Shi (2025). Model uncertainty in the cross-section of stock returns. *Journal of Econometrics*, forthcoming.

Jagannathan, R. and Y. Wang (2007). Lazy investors, discretionary consumption, and the cross-section of stock returns. *Journal of Finance 62*(4), pp. 1623–1661.

Jegadeesh, N. and S. Titman (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *Journal of Finance 48*(1), 65–91.

Jenson, M., M. Scholes, and F. Black (1972). The capital asset pricing model: Some empirical tests. *Jensen, Michael. Studies in the Theory of Capital Markets*.

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

Kan, R. and C. Zhang (1999a). GMM tests of stochastic discount factor models with useless factors. *Journal of Financial Economics 54*(1), 103–127.

Kan, R. and C. Zhang (1999b). Two-pass tests of asset pricing models with useless factors. *Journal of Finance 54*(1), 203–235.

Kelly, B. T., S. Pruitt, and Y. Su (2019). Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics 134*(3), 501–524.

Kleibergen, F. (2009). Tests of risk premia in linear factor models. *Journal of Econometrics 149*(2), 149–173.

Kleibergen, F. and Z. Zhan (2015). Unexplained factors and their effects on second pass r-squared's. *Journal of Econometrics 189*(1), 101–116.

Kleibergen, F. and Z. Zhan (2020). Robust inference for consumption-based asset pricing. *Journal of Finance 75*(1), 507–550.

Kozak, S., S. Nagel, and S. Santosh (2018). Interpreting factor models. *Journal of Finance 73*(3), 1183–1223.

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

Leibniz, G. W. (1686). *Generales Inquisitiones de Analysi Notionum et Veritatum*. Published posthumously; reproduced in *Sämtliche Schriften und Briefe*, Series VI, Vol. 4, Berlin: Akademie Verlag.

Lettau, M. and M. Pelger (2020). Factors that fit the time series and cross-section of stock returns. *Review of Financial Studies 33*(5), 2274–2325.

Lin, X. and L. Zhang (2013). The investment manifesto. *Journal of Monetary Economics 60*(3), 351–366.

Lindley, D. V. (1957). A statistical paradox. *Biometrika 44*(1–2), 187–192.

Müller, U. K. (2013). Risk of bayesian inference in misspecified models, and the sandwich covariance matrix. *Econometrica 81*(5), 1805–1849.

Newey, W. K. and K. D. West (1987). A simple, positive semidefinite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica 55*, 703–08.

Ortu, F., A. Tamoni, and C. Tebaldi (2013). Long-run risk and the persistence of consumption shocks. *Review of Financial Studies 26*(11), 2876–2915.

Parker, J. A. and C. Julliard (2005, February). Consumption risk and the cross-section of expected returns. *Journal of Political Economy 113*(1), 185–222.

Pástor, L. (2000). Portfolio selection and asset pricing models. *Journal of Finance 55*(1), 179–223.

Pástor, L. and R. F. Stambaugh (2000). Comparing asset pricing models: An investment perspective. *Journal of Financial Economics 56*(3), 335–381.

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

Pati, D., A. Bhattacharya, N. S. Pillai, and D. Dunson (2014). Posterior contraction in sparse bayesian factor models for massive covariance matrices. *Annals of Statistics 42*(3), 1102–1130.

Pukthuanthong, K., R. Roll, and A. Subrahmanyam (2019). A protocol for factor identification. *Review of Financial Studies 32*(4), 1573–1607.

Ross, S. A. (1976). The arbitrage theory of capital asset pricing. *Journal of Economic Theory 13*(3), 341–360.

Santos, T. and P. Veronesi (2022). Leverage. *Journal of Financial Economics 145*(2), 362–386.

Servin, B. and M. Stephens (2007). Imputation-based analysis of association studies: candidate regions and

quantitative traits. *PLoS Genetics 3*(7), e114.

Shanken, J. (1992). On the estimation of beta-pricing models. *Review of Financial Studies 5*(1), 1–33.

Stambaugh, R. F. and Y. Yuan (2017). Mispricing factors. *Review of Financial Studies 30*(4), 1270–1315.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 62*(4), 795–809.

Wang, G., A. Sarkar, P. Carbonetto, and M. Stephens (2020, 07). A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society Series B: Statistical Methodology 82*(5), 1273–1300.

Zhang, L. (2005). The value premium. *Journal of Finance 60*(1), 67–103.

Appendix

A Variational Inference in Cross-Sectional Pricing

In this Appendix, we illustrate how to approximate the posterior distribution $p(\boldsymbol{\lambda}, \sigma_\alpha^2 \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ of the cross-sectional regression via variational inference. The cross-sectional pricing restriction implies that

$$\boldsymbol{\mu}_r \sim \mathcal{N}\left(\boldsymbol{C}\boldsymbol{\lambda},\, \sigma_\alpha^2 \boldsymbol{I}_n\right), \qquad \boldsymbol{C} := [\mathbf{1}_n,\, \boldsymbol{\Sigma}_{rf}]\,. \tag{A1}$$

The key prior specification for risk prices and the variance of pricing errors is given by

$$\boldsymbol{\lambda} = \sum_{\ell=1}^{L} \boldsymbol{\lambda}_\ell, \quad \boldsymbol{\lambda}_\ell = \boldsymbol{\gamma}_\ell \cdot \lambda_\ell, \quad \boldsymbol{\gamma}_\ell \overset{\text{iid}}{\sim} \text{Mult}(1, \boldsymbol{\pi}), \quad \lambda_\ell \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_0^2), \quad \text{and } \pi(\sigma_\alpha^2) \propto \sigma_\alpha^{-2}, \tag{A2}$$

where the hyper-parameter σ_0^2, which reflects the prior belief about the maximal Sharpe ratios, and $\boldsymbol{\pi}$, which is set to $\frac{1}{1+p}\mathbf{1}_{p+1}$ by default, are treated as given.

A.1 Generic algorithmic design

We now take a brief detour to illustrate the rationale behind variational inference under some generic arguments. In the cross-sectional step, we let the observed data $\boldsymbol{y} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and the latent variable $\boldsymbol{z} = \left\{\sigma_\alpha^2,\, \{\boldsymbol{\gamma}_\ell,\, \lambda_\ell\}_{\ell=1}^L\right\}$. Notice that

$$\begin{aligned}
\log p(\boldsymbol{y}) &= \log \int_{\boldsymbol{z}} p(\boldsymbol{y}, \boldsymbol{z})\, \mathrm{d}\boldsymbol{z} = \log \int_{\boldsymbol{z}} \frac{p(\boldsymbol{y}, \boldsymbol{z})}{\nu(\boldsymbol{z})} \nu(\boldsymbol{z})\, \mathrm{d}\boldsymbol{z} = \log \mathbb{E}_\nu \left[\frac{p(\boldsymbol{y}, \boldsymbol{z})}{\nu(\boldsymbol{z})}\right] \\
&\geq \mathbb{E}_\nu \left[\log \frac{p(\boldsymbol{y}, \boldsymbol{z})}{\nu(\boldsymbol{z})}\right] := \text{ELBO}(\nu, \boldsymbol{y}),
\end{aligned}$$

for *any* distribution $\nu(\boldsymbol{z})$ over the latent variables $\boldsymbol{z}$. (The inequality is a direct consequence of Jensen's inequality.) $\text{ELBO}(\nu,\, \boldsymbol{y})$ is called the evidence lower bound on the log-likelihood of the observed data. The equality above holds when $p(\boldsymbol{y}, \boldsymbol{z})/\nu(\boldsymbol{z})$ does not depend on $\boldsymbol{z}$, that is, when $\nu(\boldsymbol{z}) = p(\boldsymbol{z} \mid \boldsymbol{y})$. Thus,

$$\log p(\boldsymbol{y}) = \max_{\nu} \text{ELBO}(\nu,\, \boldsymbol{y}) \geq \max_{\nu \in \mathbb{Q}} \text{ELBO}(\nu,\, \boldsymbol{y}), \tag{A3}$$

where $\mathbb{Q}$ is some *restricted* family of distributions over $\boldsymbol{z}$. The most common specification of $\mathbb{Q}$ assumes that it factorizes into independent components. That is, for $\boldsymbol{z} = \{z_1, \ldots, z_j, \ldots\}$, we consider $\nu \in \mathbb{Q}$ in equation (A3) of the form

$$\nu(\boldsymbol{z}) = \prod_j \nu_j(z_j). \tag{A4}$$

These observations motivate the variational (Bayesian) inference approach. Under the factorization in equation (A4), the constrained optimization problem over $\nu \in \mathbb{Q}$ in equation (A3) yields the following first-order condition (see the review article on variational inference by Blei, Kucukelbir, and McAuliffe (2017) for detailed derivations): for $j = 1, 2, \ldots,$

$$\log \nu_j^\star(z_j) = \mathbb{E}_{\nu_{-j}}[\log p(\boldsymbol{z}, \boldsymbol{y})] + \text{const.}, \tag{A5}$$

where $\nu_{-j}(\cdot) = \prod_{j' \neq j} \nu_{j'}(z_{j'})$ denotes the joint distribution of all variables other than z_j under the specification (A4). If we assume certain parametric forms for the marginals ν_j, and iteratively update these parameters following equation (A5) until convergence, we can solve the maximization problem in equation (A3). It is worth noting that, although the latent variables are assumed to be independent under the variational distribution and each marginal is optimized separately, they are still coupled through the update equation (A5). In fact, each latent variable z_j is optimized holding the expected values (and possibly higher-order terms, if they appear in $p(\boldsymbol{z}, \boldsymbol{y})$) of other variables fixed. Thus, the factorization in equation (A4) is also referred to as the mean-field approximation.

A.2 Variational inference for the general multi-identity model

Now we derive the variational inference algorithms for the model defined by (A1)–(A2). The full log distribution is:

$$\begin{aligned}\log p(\boldsymbol{y}, \boldsymbol{z}) = &-\frac{n}{2}\log\sigma_\alpha^2 - \frac{1}{2\sigma_\alpha^2}\left\|\boldsymbol{\mu}_r - \boldsymbol{C}\boldsymbol{\lambda}\right\|^2 + \sum_{\ell=1}^{L}\sum_{j=0}^{p}\gamma_{\ell,j}\log\pi_j - \frac{L}{2}\log\sigma_0^2 - \sum_{\ell=1}^{L}\frac{\lambda_\ell^2}{2\sigma_0^2} \\ &+ \log\pi(\sigma_\alpha^2) + \text{const.}\end{aligned} \tag{A6}$$

We consider the following mean-field approximation for the posterior of $\boldsymbol{z}$:

$$\nu(\boldsymbol{z}) = \prod_{\ell=1}^{L} \nu(\boldsymbol{\gamma}_\ell,\, \lambda_\ell)\, \nu(\sigma_\alpha^2) = \prod_{\ell=1}^{L} \nu(\lambda_\ell \mid \boldsymbol{\gamma}_\ell)\, \nu(\boldsymbol{\gamma}_\ell)\, \nu(\sigma_\alpha^2),$$

where all variational distributions ν are conditional on $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. We omit this conditioning in the notation for brevity. Specifically, we consider

$$\begin{aligned}
\nu(\lambda_\ell \mid \gamma_{\ell,j} = 1) : \lambda_\ell \mid \gamma_{\ell,j} = 1 &\sim \mathcal{N}\left(m_{\ell,j}, s_{\ell,j}^2\right), \\
\nu(\gamma_{\ell,j}) : q_{\ell,j} &= \Pr[\gamma_{\ell j} = 1], \\
\nu(\sigma_\alpha^2) : \sigma_\alpha^2 &\sim \text{Inv-Gamma}\left(\nu_\alpha,\, \sigma_{\alpha,0}^2\right),
\end{aligned}$$

and our goal is to find estimates of

$$\left\{ \{m_{\ell,j}, s_{\ell,j}^2, q_{\ell,j}\}_{\ell=1,\ldots,L; j=0,1,\ldots,p},\, \nu_\alpha,\, \sigma_{\alpha,0}^2 \right\}$$

to characterize the posterior under the mean-field approximation. To do so, we can iteratively update these parameters according to equation (A5). We begin with the key object, that is, $(\boldsymbol{\gamma}_\ell,\, \lambda_\ell)$ for all $\ell = 1, \ldots, L$.

A.2.1 Approximate posteriors of the risk prices

Update $\nu(\boldsymbol{\gamma}_\ell,\, \lambda_\ell)$ for $\ell = 1, \ldots, L$. The terms involving $(\boldsymbol{\gamma}_\ell,\, \lambda_\ell)$ are

$$-\frac{1}{2\sigma_\alpha^2}\left\|\boldsymbol{\mu}_r - \boldsymbol{C}\boldsymbol{\lambda}\right\|^2 + \sum_{\ell=1}^{L}\sum_{j=0}^{p} \gamma_{\ell,j} \log \pi_j - \sum_{\ell=1}^{L} \frac{\lambda_\ell^2}{2\sigma_0^2}.$$

Now focusing on the ℓ-th term,

$$\begin{aligned}
\log \nu^\star(\lambda_\ell \mid \gamma_{\ell,j} = 1) &= -\,\mathbb{E}_\nu\left[\frac{1}{2\sigma_\alpha^2}\right] \mathbb{E}_\nu\left\|\boldsymbol{\mu}_r - \boldsymbol{C}\boldsymbol{\lambda}_{-\ell} - \boldsymbol{C}\boldsymbol{\lambda}_\ell\right\|^2 - \frac{\lambda_\ell^2}{2\sigma_0^2} + \text{const.} \\
&= -\,\frac{1}{2\hat{\sigma}_\alpha^2}\mathbb{E}_{(\boldsymbol{\mu},\,\boldsymbol{\Sigma})}\left[-2\lambda_\ell \boldsymbol{c}_j^\top\, \boldsymbol{\mu}_{r,\ell} + \lambda_\ell^2 \|\boldsymbol{c}_j\|^2\right] - \frac{\lambda_\ell^2}{2\sigma_0^2} + \text{const.}
\end{aligned}$$

$$\nu^\star(\gamma_{\ell j} = 1) \propto \pi_j \int_\lambda \nu^\star(\lambda \mid \gamma_{\ell,j} = 1)\, \mathrm{d}\lambda \propto \pi_j \sqrt{\frac{s_{\ell,j}^2}{\sigma_0^2}} \exp\left(\frac{m_{\ell,j}^2}{2s_{\ell,j}^2}\right),$$

where $\hat{\sigma}_\alpha^2$ above is defined as follows:

$$\frac{1}{\hat{\sigma}_\alpha^2} = \mathbb{E}_\nu \left[\frac{1}{\sigma_\alpha^2} \right],$$

which will be derived explicitly later. Thus, the updating rules, conditional on $(\boldsymbol{\mu},\, \boldsymbol{\Sigma})$, are

$$\{\boldsymbol{m}_\ell,\, \boldsymbol{s}_\ell^2,\, \boldsymbol{q}_\ell, \theta_\ell\} \;\leftarrow\; \mathcal{S}(\boldsymbol{\mu}_{r,\ell},\, \boldsymbol{C},\, \hat{\sigma}_\alpha^2),$$

as defined in equation (24).

For the later presentation of our results, it is convenient to derive here the moments of the risk price vector $\boldsymbol{\lambda} = \sum_{\ell=1}^{L} \lambda_\ell \cdot \boldsymbol{\gamma}_\ell$ under the mean-field approximated distribution $\nu(\cdot)$. First, it is easy to see that

$$\bar{\boldsymbol{\lambda}} = \mathbb{E}_\nu[\boldsymbol{\lambda}] = \sum_{\ell=1}^{L} \bar{\boldsymbol{\lambda}}_\ell = \sum_{\ell=1}^{L} \boldsymbol{m}_\ell \odot \boldsymbol{q}_\ell. \tag{A7}$$

Also, we have

$$\mathbb{E}_\nu[\boldsymbol{\lambda}_\ell \boldsymbol{\lambda}_\ell^\top] = \mathbb{E}_\nu[\lambda_\ell^2 \boldsymbol{\gamma}_\ell^2] = \mathbb{E}_\nu[\lambda_\ell^2 \boldsymbol{\gamma}_\ell] = \mathrm{diag}\{(\boldsymbol{s}_\ell^2 + \boldsymbol{m}_\ell^2) \odot \boldsymbol{q}_\ell\},$$

$$\mathbb{E}_\nu[\boldsymbol{\lambda}_\ell \boldsymbol{\lambda}_{\ell'}^\top] = \mathbb{E}_\nu[\boldsymbol{\lambda}_\ell]\mathbb{E}_\nu[\boldsymbol{\lambda}_{\ell'}^\top] = \bar{\boldsymbol{\lambda}}_\ell \bar{\boldsymbol{\lambda}}_{\ell'}^\top.$$

Thus,

$$\begin{aligned}
\mathbb{E}_\nu[\boldsymbol{\lambda}\boldsymbol{\lambda}^\top] &= \sum_{\ell=1}^{L} \mathrm{diag}\{(\boldsymbol{s}_\ell^2 + \boldsymbol{m}_\ell^2) \odot \boldsymbol{q}_\ell\} + \sum_{\ell \neq \ell'} \bar{\boldsymbol{\lambda}}_\ell \bar{\boldsymbol{\lambda}}_{\ell'}^\top \\
&= \bar{\boldsymbol{\lambda}}\bar{\boldsymbol{\lambda}}^\top + \sum_{\ell=1}^{L} \left[\mathrm{diag}\{(\boldsymbol{s}_\ell^2 + \boldsymbol{m}_\ell^2) \odot \boldsymbol{q}_\ell\} - \bar{\boldsymbol{\lambda}}_\ell \bar{\boldsymbol{\lambda}}_\ell^\top \right],
\end{aligned}$$

and we have

$$\boldsymbol{V}_\lambda = \mathrm{Var}_\nu[\boldsymbol{\lambda}] = \mathbb{E}_\nu[\boldsymbol{\lambda}\boldsymbol{\lambda}^\top] - \bar{\boldsymbol{\lambda}}\bar{\boldsymbol{\lambda}}^\top = \sum_{\ell=1}^{L} \left[\mathrm{diag}\{(\boldsymbol{s}_\ell^2 + \boldsymbol{m}_\ell^2) \odot \boldsymbol{q}_\ell\} - \bar{\boldsymbol{\lambda}}_\ell \bar{\boldsymbol{\lambda}}_\ell^\top \right]. \tag{A8}$$

A.2.2 Approximate posteriors of the variance of pricing errors

Update $\nu(\sigma_\alpha^2)$. Now we turn our attention to the variance of pricing errors. The terms involving σ_α^2 are

$$-\frac{n}{2}\log\sigma_\alpha^2 - \frac{1}{2\sigma_\alpha^2}\left\|\boldsymbol{\mu}_r - \boldsymbol{C}\boldsymbol{\lambda}\right\|^2 + \log\pi(\sigma_\alpha^2).$$

Since $\pi(\sigma_\alpha^2) \propto \sigma_\alpha^{-2}$,

$$\log\nu^\star(\sigma_\alpha^2) = -\left(\frac{n}{2}+1\right)\log\sigma_\alpha^2 - \frac{1}{2\sigma_\alpha^2}\mathbb{E}_\nu\left\|\boldsymbol{\mu}_r - \boldsymbol{C}\boldsymbol{\lambda}\right\|^2 + \text{const.}$$

Thus,

$$\nu_\alpha = \frac{n}{2}, \qquad \sigma_{\alpha,0}^2 = \frac{1}{2}\mathbb{E}_\nu\left\|\boldsymbol{\mu}_r - \boldsymbol{C}\boldsymbol{\lambda}\right\|^2.$$

And according to the definition of $\hat{\sigma}_\alpha^2$ above,

$$\hat{\sigma}_\alpha^2 = \frac{1}{n}\mathbb{E}_\nu\left\|\boldsymbol{\mu}_r - \boldsymbol{C}\boldsymbol{\lambda}\right\|^2 = \frac{1}{n}\mathbb{E}_{(\boldsymbol{\mu},\,\boldsymbol{\Sigma})}\left[\left\|\boldsymbol{\mu}_r - \boldsymbol{C}\bar{\boldsymbol{\lambda}}\right\|^2 + \operatorname{tr}\left(\boldsymbol{C}^\top\boldsymbol{C}\boldsymbol{V}_\lambda\right)\right],$$

where $\bar{\boldsymbol{\lambda}}$ and $\boldsymbol{V}_\lambda$ are defined in equations (A7) and (A8).

B Posteriors of Time-Series Moments

B.1 The simple benchmark prior

Under the Jeffreys prior $\pi(\boldsymbol{\mu},\,\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-\frac{n'+1}{2}}$, we have the standard result of

$$\bar{\boldsymbol{y}} = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{y}_t, \quad \widehat{\boldsymbol{\Sigma}}_y = \frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{y}_t - \bar{\boldsymbol{y}})(\boldsymbol{y}_t - \bar{\boldsymbol{y}})^\top,$$

$$\boldsymbol{\mu} \mid \boldsymbol{\Sigma}, \{\boldsymbol{y}_t\}_{t=1}^T \sim \mathcal{N}\left(\bar{\boldsymbol{y}},\,\frac{1}{T}\boldsymbol{\Sigma}\right), \quad \text{and} \quad \boldsymbol{\Sigma} \mid \{\boldsymbol{y}_t\}_{t=1}^T \sim \text{Inv-Wishart}(T-1,\,\widehat{\boldsymbol{\Sigma}}_y).$$

B.2 The factor-based prior

Under the factor-based prior defined in (26)–(27), we do not have an analytical form for $p(\boldsymbol{\mu},\,\boldsymbol{\Sigma} \mid \{\boldsymbol{y}_t\}_{t=1}^T)$. However, we can sample $(\boldsymbol{\mu},\,\boldsymbol{\Sigma})$ using a Gibbs sampler. Specifically, we

introduce the latent variables $\{\boldsymbol{g}_t\}_{t=1}^T$ with $\boldsymbol{g}_t \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{v}, \boldsymbol{V})$, $t = 1, \ldots, T$, so that

$$\boldsymbol{y}_t \mid \boldsymbol{g}_t, \boldsymbol{u}, \boldsymbol{B}, \boldsymbol{U} \sim \mathcal{N}(\boldsymbol{u} + \boldsymbol{B}\boldsymbol{g}_t, \boldsymbol{U}). \tag{A9}$$

With the priors listed in (27), we can show that the full conditionals are

$$\begin{aligned}
&\prod_{t=1}^{T}(2\pi)^{-\frac{n'}{2}}|\boldsymbol{U}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\boldsymbol{y}_t-\boldsymbol{u}-\boldsymbol{B}\boldsymbol{g}_t)^\top\boldsymbol{U}^{-1}(\boldsymbol{y}_t-\boldsymbol{u}-\boldsymbol{B}\boldsymbol{g}_t)\right\}\\
\times&\prod_{t=1}^{T}(2\pi)^{-\frac{d}{2}}|\boldsymbol{V}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\boldsymbol{g}_t-\boldsymbol{v})^\top\boldsymbol{V}^{-1}(\boldsymbol{g}_t-\boldsymbol{v})\right\}\times|\boldsymbol{V}|^{-\frac{d+1}{2}}\times|\boldsymbol{U}|^{-\frac{d+1}{2}}.
\end{aligned}$$

To facilitate the derivation, we introduce several matrix notations, as follows:

$$\boldsymbol{Y}=\begin{pmatrix}\boldsymbol{y}_1^\top\\ \vdots\\ \boldsymbol{y}_T^\top\end{pmatrix},\quad \boldsymbol{G}=\begin{pmatrix}1 & \boldsymbol{g}_1^\top\\ \vdots & \vdots\\ 1 & \boldsymbol{g}_T^\top\end{pmatrix},\quad \boldsymbol{A}=\begin{pmatrix}\boldsymbol{u}^\top\\ \boldsymbol{B}^\top\end{pmatrix}.$$

Proposition B.1 (Gibbs sampler of the time-series parameters)**.** *Under the factor-based prior defined in equations (26)–(27) and the distributional assumption in equation (A9), the posterior distribution of* $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*, conditional on the time-series data* $\{\boldsymbol{y}_t\}_{t=1}^T$*, is given by the following conditional distributions:*

(1) Conditional on $\{\boldsymbol{y}_t\}_{t=1}^T$ *and the latent states* $\{\boldsymbol{g}_t\}_{t=1}^T$*, we update* $(\boldsymbol{u}, \boldsymbol{B}, \boldsymbol{U})$ *using a normal-inverse-Gamma distribution, as follows:*

$$\sigma_{u,i}^2 \mid \{\boldsymbol{y}_t\}_{t=1}^T, \{\boldsymbol{g}_t\}_{t=1}^T \sim \textit{Inv-Gamma}\left(\frac{T}{2}-1,\ \frac{(\boldsymbol{Y}_i-\boldsymbol{G}\widehat{\boldsymbol{A}}_i)^\top(\boldsymbol{Y}_i-\boldsymbol{G}\widehat{\boldsymbol{A}}_i)}{2}\right), \tag{A10}$$

$$\boldsymbol{A} \mid \{\boldsymbol{y}_t\}_{t=1}^T, \{\boldsymbol{g}_t\}_{t=1}^T, \boldsymbol{U} \sim \mathcal{MN}\left((\boldsymbol{G}^\top\boldsymbol{G}+\boldsymbol{D}_g)^{-1}\boldsymbol{G}^\top\boldsymbol{Y},\ \boldsymbol{U}\otimes(\boldsymbol{G}^\top\boldsymbol{G}+\boldsymbol{D}_g)^{-1}\right), \tag{A11}$$

where $\sigma_{u,i}^2$ *is the* i*-th diagonal element in* $\boldsymbol{U}$*;* $\boldsymbol{Y}_i$ *and* $\widehat{\boldsymbol{A}}_i$ *denote the* i*-th column in* $\boldsymbol{Y}$ *and* $\widehat{\boldsymbol{A}} = (\boldsymbol{G}^\top\boldsymbol{G})^{-1}\boldsymbol{G}^\top\boldsymbol{Y}$*, respectively;* $\boldsymbol{D}_g$*, which is a diagonal matrix diag*$\{0, 1, \ldots, 1\}$*, is included to avoid numerical instability during implementation.*

(2) Conditional on $\{\boldsymbol{y}_t\}_{t=1}^T$ *and* $(\boldsymbol{u}, \boldsymbol{B}, \boldsymbol{U})$*, we update both latent states* $\{\boldsymbol{g}_t\}_{t=1}^T$ *and their*

mean and covariance parameters, as follows:

$$\boldsymbol{g}_t \mid \boldsymbol{y}_t, \boldsymbol{v}, \boldsymbol{V}, \boldsymbol{u}, \boldsymbol{B}, \boldsymbol{U} \sim \mathcal{N}\Big(\Big(\boldsymbol{V}^{-1} + \boldsymbol{B}^\top \boldsymbol{U}^{-1}\boldsymbol{B}\Big)^{-1} \Big[\boldsymbol{V}^{-1}\boldsymbol{v} + \boldsymbol{B}^\top \boldsymbol{U}^{-1}(\boldsymbol{y}_t - \boldsymbol{u})\Big], \Big(\boldsymbol{V}^{-1} + \boldsymbol{B}^\top \boldsymbol{U}^{-1}\boldsymbol{B}\Big)^{-1} \Big), \tag{A12}$$

$$\boldsymbol{V} \mid \{\boldsymbol{g}_t\}_{t=1}^T \sim \text{Inv-Wishart}\left(T-1,\ \sum_{t=1}^T (\boldsymbol{g}_t - \bar{\boldsymbol{g}})(\boldsymbol{g}_t - \bar{\boldsymbol{g}})^\top\right), \tag{A13}$$

$$\boldsymbol{v} \mid \boldsymbol{V}, \{\boldsymbol{g}_t\}_{t=1}^T \sim \mathcal{N}\left(\frac{1}{T}\sum_{t=1}^T \boldsymbol{g}_t,\ \frac{1}{T}\boldsymbol{V}\right). \tag{A14}$$

(3) For each posterior draw from steps (1) and (2), we update $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *so that* $\boldsymbol{\mu} = \boldsymbol{B}\boldsymbol{v} + \boldsymbol{u}$ *and* $\boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{V}\boldsymbol{B}^\top + \boldsymbol{U}$.

First, we derive the posterior of $(\boldsymbol{u}, \boldsymbol{B}, \boldsymbol{U})$:

$$p(\boldsymbol{u}, \boldsymbol{B}, \boldsymbol{U} \mid -) \propto |\boldsymbol{U}|^{-\frac{T+d+1}{2}} \exp\left\{ -\frac{1}{2}\,\text{trace}\left[\boldsymbol{U}^{-1}(\boldsymbol{Y} - \boldsymbol{G}\boldsymbol{A})^\top(\boldsymbol{Y} - \boldsymbol{G}\boldsymbol{A})\right]\right\},$$

where $p(\cdot \mid -)$ denotes the posterior conditional on other parameters and the observed data. Define the OLS estimates of $\boldsymbol{A}$ as $\widehat{\boldsymbol{A}} = (\boldsymbol{G}^\top\boldsymbol{G})^{-1}\boldsymbol{G}^\top\boldsymbol{Y}$. We rewrite $(\boldsymbol{Y} - \boldsymbol{G}\boldsymbol{A})^\top(\boldsymbol{Y} - \boldsymbol{G}\boldsymbol{A})$ as follows:

$$(\boldsymbol{Y} - \boldsymbol{G}\boldsymbol{A})^\top(\boldsymbol{Y} - \boldsymbol{G}\boldsymbol{A}) = (\boldsymbol{Y} - \boldsymbol{G}\widehat{\boldsymbol{A}})^\top(\boldsymbol{Y} - \boldsymbol{G}\widehat{\boldsymbol{A}}) + (\boldsymbol{A} - \widehat{\boldsymbol{A}})^\top\boldsymbol{G}^\top\boldsymbol{G}(\boldsymbol{A} - \widehat{\boldsymbol{A}})$$

We sample $(\boldsymbol{u}, \boldsymbol{B})$ conditional on $\boldsymbol{U}$, the latent states, and the time-series data:

$$p(\boldsymbol{u}, \boldsymbol{B} \mid \boldsymbol{U}, -) \propto \exp\left\{ -\frac{1}{2}\,\text{trace}\left[\boldsymbol{U}^{-1}(\boldsymbol{A} - \widehat{\boldsymbol{A}})^\top\boldsymbol{G}^\top\boldsymbol{G}(\boldsymbol{A} - \widehat{\boldsymbol{A}})\right]\right\},$$

and hence, $(\boldsymbol{u}, \boldsymbol{B})$ follow a matrix normal distribution:

$$\boldsymbol{A} \mid \boldsymbol{U}, - \sim \mathcal{MN}\Big(\widehat{\boldsymbol{A}},\ \boldsymbol{U} \otimes (\boldsymbol{G}^\top\boldsymbol{G})^{-1}\Big).$$

The above posterior of $\boldsymbol{A}$ can sometimes suffer from numerical instability. To avoid this issue during implementation, we include a small L_2 penalty to $\widehat{\boldsymbol{A}}$. That is, we draw $\boldsymbol{A}$ from

a modified matrix normal distribution $\mathcal{MN}\Big((\boldsymbol{G}^\top\boldsymbol{G}+\boldsymbol{D}_g)^{-1}\boldsymbol{G}^\top\boldsymbol{Y},\ \boldsymbol{U}\otimes(\boldsymbol{G}^\top\boldsymbol{G}+\boldsymbol{D}_g)^{-1}\Big)$, where $\boldsymbol{D}_g=\text{diag}\{0,1,\ldots,1\}$.

One concern is that, if the i-th element of $\boldsymbol{y}_t$ is highly persistent, the posterior covariance $\sigma^2_{u,i}(\boldsymbol{G}^\top\boldsymbol{G}+\boldsymbol{D}_g)^{-1}$ may understate the uncertainty in estimating how y_{it} loads on the latent states, thereby distorting the posterior distribution of the factor covariance. To mitigate this source of misspecification, we follow Müller (2013) and replace this posterior covariance with its Newey–West (Newey and West, 1987) sandwich counterpart. In particular, whenever we analyze 12-quarter cumulative growth rates of macroeconomic variables—such as those reported in Tables 9, 10, 11, A8, A10, and A11—we incorporate this adjustment into the posterior distribution.

Next, we marginalise over $(\boldsymbol{u},\boldsymbol{B})$ from $p(\boldsymbol{u},\boldsymbol{B},\boldsymbol{U}\mid -)$:

$$\begin{aligned}\int p(\boldsymbol{u},\boldsymbol{B},\boldsymbol{U}\mid -)d\boldsymbol{A} &\propto \int |\boldsymbol{U}|^{-\frac{T+d+1}{2}}\exp\Big\{-\frac{1}{2}\text{trace}\left[\boldsymbol{U}^{-1}(\boldsymbol{Y}-\boldsymbol{G}\boldsymbol{A})^\top(\boldsymbol{Y}-\boldsymbol{G}\boldsymbol{A})\right]\Big\}d\boldsymbol{A}\\ &= |\boldsymbol{U}|^{-\frac{T+d+1}{2}}\exp\Big\{-\frac{1}{2}\text{trace}\left[\boldsymbol{U}^{-1}(\boldsymbol{Y}-\boldsymbol{G}\widehat{\boldsymbol{A}})^\top(\boldsymbol{Y}-\boldsymbol{G}\widehat{\boldsymbol{A}})\right]\Big\}\\ &\quad\int\exp\Big\{-\frac{1}{2}\text{trace}\left[\boldsymbol{U}^{-1}(\boldsymbol{A}-\widehat{\boldsymbol{A}})^\top\boldsymbol{G}^\top\boldsymbol{G}(\boldsymbol{A}-\widehat{\boldsymbol{A}})\right]\Big\}d\boldsymbol{A}\\ &\propto |\boldsymbol{U}|^{-\frac{T}{2}}\exp\Big\{-\frac{1}{2}\text{trace}\left[\boldsymbol{U}^{-1}(\boldsymbol{Y}-\boldsymbol{G}\widehat{\boldsymbol{A}})^\top(\boldsymbol{Y}-\boldsymbol{G}\widehat{\boldsymbol{A}})\right]\Big\}.\end{aligned}$$

Since $\boldsymbol{U}$ is a diagonal matrix, we can update each of its diagonal elements individually. In particular, the i-th diagonal element in $\boldsymbol{U}$ follows

$$p(\sigma^2_{u,i}\mid -)\propto(\sigma^2_{u,i})^{-\frac{T}{2}}\exp\Big\{-\frac{1}{2\sigma^2_{u,i}}(\boldsymbol{Y}_i-\boldsymbol{G}\widehat{\boldsymbol{A}}_i)^\top(\boldsymbol{Y}_i-\boldsymbol{G}\widehat{\boldsymbol{A}}_i)\Big\},$$

where $\boldsymbol{Y}_i$ and $\widehat{\boldsymbol{A}}_i$ denote the i-th column in $\boldsymbol{Y}$ and $\widehat{\boldsymbol{A}}$, respectively. The above derivation implies that we can sample $\{\sigma^2_{u,i}\}_{i=1}^{n'}$ from inverse-gamma distributions as in equation (A10).

We now derive the posterior of the latent states $\{\boldsymbol{g}_t\}_{t=1}^T$. For each time t,

$$\begin{aligned}p(\boldsymbol{g}_t\mid -)&\propto\exp\Big\{-\frac{1}{2}(\boldsymbol{y}_t-\boldsymbol{u}-\boldsymbol{B}\boldsymbol{g}_t)^\top\boldsymbol{U}^{-1}(\boldsymbol{y}_t-\boldsymbol{u}-\boldsymbol{B}\boldsymbol{g}_t)-\frac{1}{2}(\boldsymbol{g}_t-\boldsymbol{v})^\top\boldsymbol{V}^{-1}(\boldsymbol{g}_t-\boldsymbol{v})\Big\}\\ &\propto\exp\Big\{-\frac{1}{2}\boldsymbol{g}_t^\top\left(\boldsymbol{V}^{-1}+\boldsymbol{B}^\top\boldsymbol{U}^{-1}\boldsymbol{B}\right)\boldsymbol{g}_t+\boldsymbol{g}_t^\top\left[\boldsymbol{V}^{-1}\boldsymbol{v}+\boldsymbol{B}^\top\boldsymbol{U}^{-1}(\boldsymbol{y}_t-\boldsymbol{u})\right]\Big\};\end{aligned}$$

therefore, $\boldsymbol{g}_t$ follows a multivariate normal posterior distribution as in equation (A12).

Finally, we derive the posterior of $(\boldsymbol{v}, \boldsymbol{V})$. Conditional on other parameters and time-series data, we can show that

$$p(\boldsymbol{v}, \boldsymbol{V} \mid -) \propto |\boldsymbol{V}|^{-\frac{T+d+1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{g}_t - \boldsymbol{v})^\top \boldsymbol{V}^{-1}(\boldsymbol{g}_t - \boldsymbol{v})\right\},$$

which is the kernel of the Normal–inverse–Wishart conjugate in equations (A13)–(A14).

When we use a large value of d in the estimation, we need to perform an additional transformation at each step. In particular, we transform $\boldsymbol{g}_t$ into uncorrelated latent states; that is, we construct $\boldsymbol{H}\boldsymbol{g}_t$ such that $\text{Cov}(\boldsymbol{H}\boldsymbol{g}_t) = \boldsymbol{I}_d$ and then use the transformed latent states $\tilde{\boldsymbol{g}}_t = \boldsymbol{H}\boldsymbol{g}_t$ in the next sampling step. To see why this transformation does not affect the estimates of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we rewrite equations (26) and (A9) as follows:

$$\boldsymbol{\mu} = \boldsymbol{u} + \underbrace{\boldsymbol{B}\boldsymbol{H}^{-1}}_{\tilde{\boldsymbol{B}}}\underbrace{\boldsymbol{H}\boldsymbol{v}}_{\tilde{\boldsymbol{v}}}, \quad \boldsymbol{\Sigma} = \underbrace{\boldsymbol{B}\boldsymbol{H}^{-1}}_{\tilde{\boldsymbol{B}}}\underbrace{\boldsymbol{H}\boldsymbol{V}\boldsymbol{H}^\top}_{\tilde{\boldsymbol{V}}}\underbrace{(\boldsymbol{B}\boldsymbol{H}^{-1})^\top}_{\tilde{\boldsymbol{B}}^\top} + \boldsymbol{U},$$

$$\boldsymbol{y}_t \mid \tilde{\boldsymbol{g}}_t, \boldsymbol{u}, \tilde{\boldsymbol{B}}, \boldsymbol{U} \sim \mathcal{N}(\boldsymbol{u} + \tilde{\boldsymbol{B}}\tilde{\boldsymbol{g}}_t,\, \boldsymbol{U}), \text{ and } \tilde{\boldsymbol{g}}_t \sim \mathcal{N}(\tilde{\boldsymbol{v}}, \tilde{\boldsymbol{V}}).$$

Thanks to the above rotation invariance property, this linear transformation does not affect the estimates (and the posterior distribution) of mean returns and the covariance matrix. However, this transformation is needed to avoid occasional instability in the numerical implementation.

C Proofs

C.1 Proof of Proposition 1

Proof. It is straightforward to show that

$$\text{Var}_i[\hat{c}_i] = V_c + \frac{\sigma_e^2}{T} + O_p(T^{-1/2})$$

$$\text{Var}_i[\hat{\tilde{c}}_i] = \rho^2 V_c + \frac{\sigma_e^2}{T} + O_p(T^{-1/2})$$

$$\mathrm{Cov}_i(\hat{c}_i, \hat{\tilde{c}}_i) = \rho V_c + \frac{\rho\sigma_e^2}{T} + O_p(T^{-1/2})$$

Then we have,

$$\mathrm{Corr}_i(\hat{c}_i, \hat{\tilde{c}}_i) = \frac{\rho V_c + \frac{\rho\sigma_e^2}{T} + O_p(T^{-1/2})}{\sqrt{\left(V_c + \frac{\sigma_e^2}{T} + O_p(T^{-1/2})\right)\left(\rho^2 V_c + \frac{\sigma_e^2}{T} + O_p(T^{-1/2})\right)}}. \tag{A15}$$

Since $V_c > 0$ and $\rho \neq 0$, the denominator is bounded away from 0 w.p.a.1 and

$$\sqrt{\left(V_c + \frac{\sigma_e^2}{T} + O_p(T^{-1/2})\right)\left(\rho^2 V_c + \frac{\sigma_e^2}{T} + O_p(T^{-1/2})\right)} = |\rho| V_c + O_p(T^{-1/2}).$$

Also the numerator satisfies

$$\rho V_c + \frac{\rho\sigma_e^2}{T} + O_p(T^{-1/2}) = \rho V_c + O_p(T^{-1/2}),$$

Therefore,

$$\mathrm{Corr}_i(\hat{c}_i, \hat{\tilde{c}}_i) = \frac{\rho V_c + O_p(T^{-1/2})}{|\rho| V_c + O_p(T^{-1/2})} = \frac{\rho}{|\rho|} + O_p(T^{-1/2}) = \mathrm{sgn}(\rho) + O_p(T^{-1/2}). \tag{A16}$$

□

C.2 Proof of Proposition 3

Proof. With elements of $\boldsymbol{c}_j$ being iid F_c with mean μ_c and variance V_c, we have, by the strong law of large numbers,

$$\frac{1}{n}\|\boldsymbol{c}_j\|^2 \xrightarrow{a.s.} q,$$

where $q = \mu_c^2 + V_c$. Now recall that $\boldsymbol{\mu}_r = \lambda^\star \boldsymbol{c}_j$, $\tilde{\boldsymbol{c}}_j = \rho \boldsymbol{c}_j$, it is straightforward to show that

$$\frac{w_j}{w_{\tilde{j}}} = \frac{\pi_j}{\pi_{\tilde{j}}} \sqrt{\frac{\sigma_0^2\rho^2\|\boldsymbol{c}_j\|^2 + \sigma_\alpha^2}{\sigma_0^2\|\boldsymbol{c}_j\|^2 + \sigma_\alpha^2}} \exp\left[\frac{(1-\rho^2)\sigma_0^2(\lambda^\star)^2\|\boldsymbol{c}_j\|^4}{2(\sigma_0^2\rho^2\|\boldsymbol{c}_j\|^2 + \sigma_\alpha^2)(\sigma_0^2\|\boldsymbol{c}_j\|^2 + \sigma_\alpha^2)}\right].$$

Consider functions

$$g(x) = \sqrt{\frac{\sigma_0^2\rho^2 x + \sigma_\alpha^2}{\sigma_0^2 x + \sigma_\alpha^2}}, \qquad h(x) = \frac{\sigma_0^4 x^2}{(\sigma_0^2\rho^2 x + \sigma_\alpha^2)(\sigma_0^2 x + \sigma_\alpha^2)},$$

we have $\lim_{x\to\infty} g(x) = |\rho|$ and $\lim_{x\to\infty} h(x) = 1/\rho^2$. Since $\|\boldsymbol{c}_j\|^2 \xrightarrow{a.s.} \infty$ as $n \to \infty$, by the continuous mapping theorem,

$$\frac{w_j}{w_{\tilde{\jmath}}} \xrightarrow{a.s.} \frac{\pi_j}{\pi_{\tilde{\jmath}}}|\rho| \exp\left[\frac{(1-\rho^2)\sigma_0^2(\lambda^\star)^2}{2}\frac{1}{\rho^2\sigma_0^4}\right] = \frac{\pi_j}{\pi_{\tilde{\jmath}}}|\rho| \exp\left[\left(\frac{1}{\rho^2} - 1\right)\frac{(\lambda^\star)^2}{2\sigma_0^2}\right].$$

By the definitions of q_j and $q_{\tilde{\jmath}}$ with respect to w_js, the first part of the proposition follows.

For the second part of the proposition, consider the function

$$f(\rho) = |\rho| \exp\left(k\left(\frac{1}{\rho^2} - 1\right)\right), \qquad k := \frac{(\lambda^\star)^2}{2\sigma_0^2}.$$

To complete the proof, we need to show that $f(\rho) \geq 1$. By the assumption that $\lambda^\star \geq \sigma_0$, $k \geq 1/2$. Taking the logarithm and derivative,

$$\ell(\rho) := \log f(\rho) = \log|\rho| + k\left(\frac{1}{\rho^2} - 1\right), \qquad \ell'(\rho) = \frac{1}{\rho} - \frac{2k}{\rho^3}, \qquad (\rho \neq 0).$$

For $\rho \in (0,\, 1]$, $\ell'(\rho) \leq 1/\rho(1 - 1/\rho^2) < 0$, thus, $f(\rho)$ is decreasing in $(0,\, 1]$ and $f(\rho) \geq f(1) = 1$. For $\rho \in [-1,\, 0)$, $\ell'(\rho) \geq 1/\rho(1 - 1/\rho^2) > 0$, thus, $f(\rho)$ is increasing in $[-1,\, 0)$ and $f(\rho) \geq f(-1) = 1$. □

D Additional Tables and Figures

Table A1: List of Equity Anomalies

Category	Test assets
value related	value, valuem, divp, ep, cfp, dur, sp, valmom, valmomprof, valprof
investment	inv, invcap, igrowth, growth, noa, gltnoa, sgrowth
profitability	prof, roaa, roea, roa, roe, rome, gmargins, aturnover
trading frictions	ivol, shvol, betaarb, sue, shortint
issuance	repurch, debtiss, ciss, nissa, nissm
momentum & reversal	mom, mom12, indmom, momrev, lrrev, strev, indmomrev, indrrev, indrrevlv
others	size, price, accruals, age, fscore, lev, season

The table presents a list of equity anomalies used as test assets in Section 3. In particular, ten decile portfolios are constructed for each of the 51 firm characteristics listed above. We include both the long and short legs (i.e., the decile 10 and decile 1 portfolios), resulting in 102 equity portfolio returns. The data are from Serhiy Kozak's website, and the sample period runs from January 1974 to December 2019.

Table A2: List of Factors

Description of factors:	Frequency	Source
Five factors in Fama and French (2015) (MKT, SMB, HML, RMW, CMA)	Monthly	Ken French's website
Momentum factor in Jegadeesh and Titman (1993) (MOM)	Monthly	Ken French's website
Hedged factors in Daniel et al. (2020) (MKT$\star$, SMB$\star$, HML$\star$, RMW$\star$, CMA$\star$)	Monthly	Kent Daniel's website
q-factors in Hou et al. (2015) (ME, IA, ROE)	Monthly	q-factor library
Expected growth factor in Hou et al. (2021) (EG)	Monthly	q-factor library
Mispricing factors in Stambaugh and Yuan (2017) (MGMT, PERF)	Monthly	Global Factor Data
Short- and long-horizon behavioral factors in Daniel et al. (2020) (FIN, PEAD)	Monthly	Kent Daniel's website
AEM intermediary factor in Adrian et al. (2014)	Quarterly	Tyler Muir's Website
HKM intermediary factor in He et al. (2017)	Monthly	Zhiguo He's website
Liquidity factor in Pástor and Stambaugh (2003)	Monthly	Lubos Pastor's website
Industrial production growth (log change in real per capita)	Quarterly	Federal Reserve Bank of St. Louis
GDP growth (log change in real per capita)	Quarterly	BEA Table 7.1
Durable, nondurable, & service consumption growth (log change in real per capita)	Quarterly	BEA Table 7.1
Total factor productivity (TFP) growth, utilization-adjusted	Quarterly	John Fernald's website
Unemployment rate change	Quarterly	Federal Reserve Bank of St. Louis
Hours worked, log growth	Quarterly	Federal Reserve Bank of St. Louis
Real Investment per capita, log growth	Quarterly	Federal Reserve Bank of St. Louis

The table presents a list of factors used in Section 3. For each variable, we show the name, description, frequency, and data source. Following Angeletos et al. (2020), hours worked are defined as the log(Nonfarm business sector: average weekly hours $\times$ Employment Level / Civilian non-institutional population); real investment per capita is defined as log[(Share of GDP: personal durable consumption expenditures + Share of GDP: gross private domestic investment) $\times$ Real gross domestic product per capita]. The sample spans 1974Q1–2019Q4, except for the AEM intermediary factor, which ends in 2017Q3.

Table A3: HKM vs. AEM Intermediary Factors, Alternative Priors

	Panel A. $\sigma_0 = 0.3/\sqrt{4}$				Panel B. $\sigma_0 = 0.5/\sqrt{4}$			
	Factor prob. within identities, $\boldsymbol{q}_\ell$		$p(\boldsymbol{m}_\ell \mid \text{data})$	$\mathbb{E}[\boldsymbol{\lambda} \mid \text{data}]$	Factor prob. within identities, $\boldsymbol{q}_\ell$		$p(\boldsymbol{m}_\ell \mid \text{data})$	$\mathbb{E}[\boldsymbol{\lambda} \mid \text{data}]$
	$\ell = 1$	$\ell = 2$	mean & 90% CI		$\ell = 1$	$\ell = 2$	mean & 90% CI	
$\mathbf{1}_n$	0.642	0	0.180 (0.043, 0.317)	0.116	0.558	0	0.175 (0.028, 0.316)	0.097
MKT	0.100	0	0.201 (0.048, 0.353)	0.020	0.213	0	0.195 (0.031, 0.353)	0.041
HKM	0.258	0	0.253 (0.061, 0.449)	0.065	0.230	0	0.245 (0.039, 0.446)	0.056
AEM	0	1.000	0.366 (0.141, 0.584)	0.366	0	1.000	0.482 (0.184, 0.775)	0.482
Ident. prob.	0.987	0.942			0.976	0.956		
$\mathbb{E}[\lambda_\ell \mid \text{data}]$	0.201	0.366			0.195	0.482		

This table presents the Bayesian estimates of two intermediary factors, HKM (He et al., 2017) and AEM (Adrian et al., 2014), controlling for a common intercept ($\mathbf{1}_n$) and the market factor (MKT). The data are at the quarterly frequency, spanning from 1974Q1 to 2017Q3. In empirical applications, all variables are standardized to have unit variance per period. The data description is in Section 3.1. We use Algorithm 1, with $L = 4$ and $\sigma_0 \in \{0.3/\sqrt{4}, 0.5/\sqrt{4}\}$, to estimate the posterior factor probabilities within identities ($\{\boldsymbol{q}_\ell\}_{\ell=1}^L$). The second last row reports the posterior identity probabilities ($\{\theta_\ell\}_{\ell=1}^L$). We display the identities with posterior probabilities above 50%. Conditional on $\{\boldsymbol{q}_\ell\}_{\ell=1}^L$, we estimate the posterior distribution of factor risk prices using the approach described in Section 1.3.2. We present both conditional and unconditional factor risk prices. In the column "$p(\boldsymbol{m}_\ell \mid \text{data})$", we report the factor risk prices conditional on the factors being selected (both the conditional posterior mean and 90% credible intervals). In the column "$\mathbb{E}[\boldsymbol{\lambda} \mid \text{data}]$", we display the unconditional posterior mean of the factor risk prices, defined as $\mathbb{E}[\sum_{\ell=1}^L \boldsymbol{m}_\ell \odot \boldsymbol{q}_\ell \mid \text{data}]$. In the row labeled "$\mathbb{E}[\lambda_\ell \mid \text{data}]$", we report the posterior mean of the identity market price of risk, defined as $\mathbb{E}[\sum_j q_{\ell,j} m_{\ell,j} \mid \text{data}]$.

Table A4: PCs vs. RP-PCs, Alternative Priors

	Panel A. $\sigma_0 = 0.3/\sqrt{12}$							Panel B. $\sigma_0 = 0.5/\sqrt{12}$							
	Factor prob. within identities, $\boldsymbol{q}_\ell$					$p(\boldsymbol{m}_\ell \mid \text{data})$	$\mathbb{E}[\boldsymbol{\lambda} \mid \text{data}]$	Factor prob. within identities, $\boldsymbol{q}_\ell$						$p(\boldsymbol{m}_\ell \mid \text{data})$	$\mathbb{E}[\boldsymbol{\lambda} \mid \text{data}]$
	$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 4$	$\ell = 5$	mean & 90% CI		$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 4$	$\ell = 5$	$l = 6$	mean & 90% CI	
$\mathbf{1}_n$	0.162	0	0	0	0	0.123 (0.058, 0.189)	0.020	0.009	0	0	0	0	0		0.001
PC1	0.398	0	0	0	0	0.137 (0.065, 0.212)	0.055	0.484	0	0	0	0	0	0.138 (0.064, 0.216)	0.067
PC2	0	0	0	0	0.108	-0.043 (-0.102, 0.016)	-0.005	0	0	0	0	0.079	0	-0.046 (-0.113, 0.019)	-0.004
PC3	0	0	0	0	0.005		0.003	0	0	0	0	0	0		0
PC4	0	0	0	0.255	0.001	0.086 (0.022, 0.150)	0.026	0	0	0	0.991	0	0	0.130 (0.048, 0.211)	0.129
PC5	0	0	0	0	0.002		-0.032	0	0	0	0	0	1.000	-0.233 (-0.348, -0.122)	-0.233
RP-PC1	0.439	0	0	0	0	0.137 (0.065, 0.212)	0.060	0.507	0	0	0	0	0	0.138 (0.064, 0.216)	0.070
RP-PC2	0	1.000	0	0	0.001	0.232 (0.163, 0.301)	0.236	0	1.000	0	0	0	0	0.251 (0.173, 0.331)	0.251
RP-PC3	0	0	0	0	0.879	0.053 (-0.015, 0.121)	0.049	0	0	0	0	0.920	0	0.055 (-0.021, 0.132)	0.050
RP-PC4	0	0	0	0.745	0.001	0.088 (0.023, 0.151)	0.069	0	0	0	0.009	0	0		0.001
RP-PC5	0	0	1.000	0	0.002	0.211 (0.156, 0.266)	0.233	0	0	1.000	0	0	0	0.460 (0.342, 0.571)	0.460
Ident. prob.	0.998	1.000	1.000	0.835	0.662			0.998	1.000	1.000	0.945	0.661	0.987		
$\mathbb{E}[\lambda_\ell \mid \text{data}]$	0.135	0.232	0.211	0.088	0.052			0.138	0.251	0.460	0.130	0.054	-0.233		

This table presents the Bayesian estimates of the first five PCs and RP-PCs (Lettau and Pelger, 2020), controlling for a common intercept ($\mathbf{1}_n$). The data are at the monthly frequency, spanning from January 1974 to December 2019. In empirical applications, all variables are standardized to have unit variance per period. The data description is in Section 3.1. We use Algorithm 1, with $L = 10$ and $\sigma_0 \in \{0.3/\sqrt{12}, 0.5/\sqrt{12}\}$, to estimate the posterior factor probabilities within identities ($\{\boldsymbol{q}_\ell\}_{\ell=1}^L$). The second last row reports the posterior identity probabilities ($\{\theta_\ell\}_{\ell=1}^L$). We display the identities with posterior probabilities above 50%. Conditional on $\{\boldsymbol{q}_\ell\}_{\ell=1}^L$, we estimate the posterior distribution of factor risk prices using the approach described in Section 1.3.2. We present both conditional and unconditional factor risk prices. In the column "$p(\boldsymbol{m}_\ell \mid \text{data})$", we report the factor risk prices conditional on the factors being selected (both the conditional posterior mean and 90% credible intervals). In the column "$\mathbb{E}[\boldsymbol{\lambda} \mid \text{data}]$", we display the unconditional posterior mean of the factor risk prices, defined as $\mathbb{E}[\sum_{\ell=1}^L \boldsymbol{m}_\ell \odot \boldsymbol{q}_\ell \mid \text{data}]$. In the row labeled "$\mathbb{E}[\lambda_\ell \mid \text{data}]$", we report the posterior mean of the identity market price of risk, defined as $\mathbb{E}[\sum_j q_{\ell,j} m_{\ell,j} \mid \text{data}]$.

Table A5: Fama-French Factors vs. Q-Factors, Alternative Priors

	Panel A. $\sigma_0 = 0.3/\sqrt{12}$						Panel B. $\sigma_0 = 0.5/\sqrt{12}$					
	Factor prob. within identities, $\boldsymbol{q}_\ell$				$p(\boldsymbol{m}_\ell \mid \text{data})$	$\mathbb{E}[\boldsymbol{\lambda} \mid \text{data}]$	Factor prob. within identities, $\boldsymbol{q}_\ell$				$p(\boldsymbol{m}_\ell \mid \text{data})$	$\mathbb{E}[\boldsymbol{\lambda} \mid \text{data}]$
	$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 4$	mean & 90% CI		$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 4$	mean & 90% CI	
$\mathbf{1}_n$	0.986	0	0.005	0	0.186 (0.118, 0.254)	0.184	0.971	0	0	0	0.191 (0.121, 0.262)	0.185
MKT	0.014	0	0.004	0	0.209 (0.132, 0.285)	0.003	0.029	0	0	0	0.215 (0.136, 0.294)	0.006
SMB	0	0	0.406	0	0.071 (-0.004, 0.147)	0.029	0	0	0.448	0	0.101 (0.010, 0.191)	0.045
HML	0	0	0	0.941	0.131 (0.054, 0.206)	0.123	0	0	0	0.965	0.143 (0.061, 0.225)	0.138
RMW	0	0	0.003	0		0	0	0	0.001	0		0
CMA	0	0	0	0.020	0.095 (0.040, 0.153)	0.002	0	0	0	0.013	0.104 (0.046, 0.164)	0.001
ME	0	0	0.581	0	0.072 (-0.004, 0.150)	0.042	0	0	0.550	0	0.103 (0.010, 0.196)	0.056
IA	0	0	0	0.038	0.103 (0.043, 0.163)	0.004	0	0	0	0.022	0.113 (0.049, 0.177)	0.002
ROE	0	1.000	0.001	0	0.230 (0.151, 0.307)	0.230	0	1.000	0	0	0.274 (0.178, 0.369)	0.274
Ident. prob.	1.000	1.000	0.839	0.982			1.000	1.000	0.905	0.983		
$\mathbb{E}[\lambda_\ell \mid \text{data}]$	0.186	0.230	0.071	0.129			0.192	0.274	0.102	0.142		

This table presents the Bayesian estimates of the Fama-French five factors (Fama and French, 2015) and q-factors (Hou et al., 2015), controlling for a common intercept ($\mathbf{1}_n$). The data are at the monthly frequency, spanning from January 1974 to December 2019. In empirical applications, all variables are standardized to have unit variance per period. The data description is in Section 3.1. We use Algorithm 1, with $L = 10$ and $\sigma_0 \in \{0.3/\sqrt{12}, 0.5/\sqrt{12}\}$, to estimate the posterior factor probabilities within identities ($\{\boldsymbol{q}_\ell\}_{\ell=1}^L$). The second last row reports the posterior identity probabilities ($\{\theta_\ell\}_{\ell=1}^L$). We display the identities with posterior probabilities above 50%. Conditional on $\{\boldsymbol{q}_\ell\}_{\ell=1}^L$, we estimate the posterior distribution of factor risk prices using the approach described in Section 1.3.2. We present both conditional and unconditional factor risk prices. In the column "$p(\boldsymbol{m}_\ell \mid \text{data})$", we report the factor risk prices conditional on the factors being selected (both the conditional posterior mean and 90% credible intervals). In the column "$\mathbb{E}[\boldsymbol{\lambda} \mid \text{data}]$", we display the unconditional posterior mean of the factor risk prices, defined as $\mathbb{E}[\sum_{\ell=1}^L \boldsymbol{m}_\ell \odot \boldsymbol{q}_\ell \mid \text{data}]$. In the row labeled "$\mathbb{E}[\lambda_\ell \mid \text{data}]$", we report the posterior mean of the identity market price of risk, defined as $\mathbb{E}[\sum_j q_{\ell,j} m_{\ell,j} \mid \text{data}]$.

Table A6: Statistical vs. Characteristic-Based Factors, Alternative Priors

	$\sigma_0 = 0.3/\sqrt{12}$						$\sigma_0 = 0.5/\sqrt{12}$						
	Factor prob. within identities, $\boldsymbol{q}_\ell$				$p(\boldsymbol{m}_\ell \mid \text{data})$	$\mathbb{E}[\boldsymbol{\lambda} \mid \text{data}]$	Factor prob. within identities, $\boldsymbol{q}_\ell$					$p(\boldsymbol{m}_\ell \mid \text{data})$	$\mathbb{E}[\boldsymbol{\lambda} \mid \text{data}]$
	$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 4$	mean & 90% CI		$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 4$	$\ell = 5$	mean & 90% CI	
Panel A													
$\mathbf{1}_n$	0.015	0	0	0	0.124 (0.059, 0.192)	0.002	0.001	0	0	0	0		0
PC1	0.188	0	0	0	0.139 (0.066, 0.214)	0.026	0.164	0	0	0	0	0.140 (0.065, 0.218)	0.023
PC2	0	0	0	0		-0.004	0	0	0	0	0		-0.006
PC3	0	0	0	0		0.005	0	0	0	0	0		0.003
PC4	0	0	0	0.736	0.091 (0.024, 0.157)	0.070	0	0	0	0.994	0	0.128 (0.046, 0.206)	0.128
PC5	0	0	0	0		-0.016	0	0	0	0	1.000	-0.211 (-0.324, -0.096)	-0.214
RP-PC1	0.212	0	0	0	0.139 (0.066, 0.214)	0.029	0.182	0	0	0	0	0.140 (0.065, 0.218)	0.025
RP-PC2	0	1.000	0	0	0.213 (0.142, 0.287)	0.216	0	1.000	0	0	0	0.239 (0.161, 0.317)	0.239
RP-PC3	0	0	0	0		0.010	0	0	0	0	0		0.013
RP-PC4	0	0	0	0.263	0.085 (0.020, 0.151)	0.025	0	0	0	0.006	0		0.001
RP-PC5	0	0	1.000	0	0.206 (0.145, 0.267)	0.222	0	0	1.000	0	0	0.430 (0.314, 0.548)	0.437
MKT	0.584	0	0	0	0.140 (0.066, 0.216)	0.082	0.654	0	0	0	0	0.141 (0.066, 0.220)	0.092
SMB	0	0	0	0		-0.001	0	0	0	0	0		-0.001
HML	0	0	0	0		0	0	0	0	0	0		0
RMW	0	0	0	0		0	0	0	0	0	0		0.001
CMA	0	0	0	0		0	0	0	0	0	0		0
MOM	0	0	0	0.001		0.006	0	0	0	0	0		0.004
ME	0	0	0	0		-0.001	0	0	0	0	0		-0.001
IA	0	0	0	0		0	0	0	0	0	0		0
ROE	0	0	0	0		0.001	0	0	0	0	0		0.002
EG	0	0	0	0		0	0	0	0	0	0		0
MGMT	0	0	0	0		0	0	0	0	0	0		0
PERF	0	0	0	0		0.004	0	0	0	0	0		0.006
PEAD	0	0	0	0		0.007	0	0	0	0	0		0.007
FIN	0	0	0	0		0	0	0	0	0	0		0
Ident. prob.	0.998	1.000	0.996	0.804			0.998	1.000	1.000	0.921	0.958		
$\mathbb{E}[\lambda_\ell \mid \text{data}]$	0.139	0.213	0.206	0.089			0.141	0.239	0.430	0.128	-0.211		
Panel B													
$\mathbf{1}_n$	0	0	0	0		0	0	0	0	0			0
PC1	0	0	0	0		0	0	0	0	0			0
PC2	0	0	0	0		0.003	0	0	0	0			0.003
PC3	0	0	0	0		-0.002	0	0	0	0			-0.002
PC4	0	0.001	0	0		0.002	0	0	0	0			0
PC5	0	0	0	0		-0.002	0	0	0	0			-0.023
RP-PC1	0	0	0	0		0	0	0	0	0			0
RP-PC2	0	0	0	1.000	0.124 (0.029, 0.215)	0.125	0	0	0	1.000		0.133 (0.031, 0.229)	0.133
RP-PC3	0	0	0	0		-0.004	0	0	0	0			-0.004
RP-PC4	0	0.999	0	0	0.155 (0.076, 0.230)	0.157	0	1.000	0	0		0.171 (0.085, 0.255)	0.172
RP-PC5	0	0	1.000	0	0.231 (0.165, 0.295)	0.251	0	0	1.000	0		0.297 (0.217, 0.373)	0.301
MKT$\star$	1.000	0	0	0	0.280 (0.131, 0.427)	0.280	1.000	0	0	0		0.285 (0.130, 0.439)	0.285
SMB$\star$	0	0	0	0		0.005	0	0	0	0			0.002
HML$\star$	0	0	0	0		0.001	0	0	0	0			0.002
RMW$\star$	0	0	0	0		-0.015	0	0	0	0			-0.017
CMA$\star$	0	0	0	0		0	0	0	0	0			0.001
MOM	0	0	0	0		-0.002	0	0	0	0			-0.002
ME	0	0	0	0		0.002	0	0	0	0			0.001
IA	0	0	0	0		0	0	0	0	0			0
ROE	0	0	0	0		-0.002	0	0	0	0			-0.002
EG	0	0	0	0		-0.001	0	0	0	0			0
MGMT	0	0	0	0		0	0	0	0	0			0
PERF	0	0	0	0		-0.006	0	0	0	0			-0.007
PEAD	0	0	0	0		-0.006	0	0	0	0			-0.005
FIN	0	0	0	0		0	0	0	0	0			0
Ident. prob.	0.998	0.978	0.999	0.914			0.998	0.981	1.000	0.913			
$\mathbb{E}[\lambda_\ell \mid \text{data}]$	0.280	0.155	0.231	0.124			0.285	0.171	0.297	0.133			

This table presents the Bayesian estimates of the statistical and characteristic-based factors. The data are at the monthly frequency, spanning from January 1974 to December 2019. In empirical applications, all variables are standardized to have unit variance per period. The data description is in Section 3.1. In Panel A, we use the Fama-French five factors (Fama and French, 2015), whereas the hedged returns (Daniel et al., 2020) of the Fama-French five factors are employed in Panel B. We use Algorithm 1, with $L = 10$ and $\sigma_0 \in \{0.3/\sqrt{12}, 0.5/\sqrt{12}\}$, to estimate the posterior factor probabilities within identities ($\{\boldsymbol{q}_\ell\}_{\ell=1}^L$). The second last row reports the posterior identity probabilities ($\{\theta_\ell\}_{\ell=1}^L$). We display the identities with posterior probabilities above 50%. Conditional on $\{\boldsymbol{q}_\ell\}_{\ell=1}^L$, we estimate the posterior distribution of factor risk prices using the approach described in Section 1.3.2. We present both conditional and unconditional factor risk prices. In the column "$p(\boldsymbol{m}_\ell \mid \text{data})$", we report the factor risk prices conditional on the factors being selected (both the conditional posterior mean and 90% credible intervals). In the column "$\mathbb{E}[\boldsymbol{\lambda} \mid \text{data}]$", we display the unconditional posterior mean of the factor risk prices, defined as $\mathbb{E}[\sum_{\ell=1}^L \boldsymbol{m}_\ell \odot \boldsymbol{q}_\ell \mid \text{data}]$. In the row labeled "$\mathbb{E}[\lambda_\ell \mid \text{data}]$", we report the posterior mean of the identity market price of risk, defined as $\mathbb{E}[\sum_j q_{\ell,j} m_{\ell,j} \mid \text{data}]$.

Table A7: Statistical vs. Characteristic-Based Factors, Excluding Risk-Premia PCs

	Panel A											
	$\sigma_0 = 0.3/\sqrt{4}$				$\sigma_0 = 0.4/\sqrt{4}$				$\sigma_0 = 0.5/\sqrt{4}$			
	$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 4$	$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 4$	$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 4$
$\mathbf{1}_n$	0.929	0	0	0	0.911	0	0	0	0.899	0	0	0
PC1	0.044	0	0	0	0.057	0	0	0	0.064	0	0	0
PC2	0	0	0	0.002	0	0	0	0	0	0	0	0
PC3	0	1.000	0	0.001	0	1.000	0	0	0	1.000	0	0
PC4	0	0	1.000	0.002	0	0	1.000	0	0	0	1.000	0
PC5	0	0	0	0.984	0	0	0	0.996	0	0	0	0.998
MKT	0.026	0	0	0	0.033	0	0	0	0.036	0	0	0
SMB	0	0	0	0.001	0	0	0	0	0	0	0	0
HML	0	0	0	0.001	0	0	0	0	0	0	0	0
RMW	0	0	0	0.001	0	0	0	0	0	0	0	0
CMA	0	0	0	0.001	0	0	0	0	0	0	0	0
MOM	0	0	0	0.001	0	0	0	0	0	0	0	0
ME	0	0	0	0.001	0	0	0	0	0	0	0	0
IA	0	0	0	0.001	0	0	0	0	0	0	0	0
ROE	0	0	0	0.001	0	0	0	0	0	0	0	0
EG	0	0	0	0.001	0	0	0	0	0	0	0	0
MGMT	0	0	0	0.001	0	0	0	0	0	0	0	0
PERF	0	0	0	0.001	0	0	0	0	0	0	0	0
PEAD	0	0	0	0.002	0	0	0	0.001	0	0	0	0
FIN	0	0	0	0	0	0	0	0	0	0	0	0
Ident. prob.	0.998	0.989	0.863	0.727	0.998	0.990	0.873	0.780	0.997	0.989	0.875	0.798
	Panel B											
	$\sigma_0 = 0.3/\sqrt{4}$				$\sigma_0 = 0.4/\sqrt{4}$				$\sigma_0 = 0.5/\sqrt{4}$			
	$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 4$	$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 4$	$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 4$
$\mathbf{1}_n$	0	0	0	0	0	0	0	0	0	0	0	0
PC1	0	0	0	0	0	0	0	0	0	0	0	0
PC2	0	0	0	0.941	0	0	0	0.955	0	0	0	0.958
PC3	0	0	0	0.001	0	0	0	0	0	0	0	0
PC4	0	1.000	0	0.001	0	1.000	0	0.001	0	1.000	0	0
PC5	0	0	1.000	0.002	0	0	1.000	0.001	0	0	1.000	0.001
MKT⋆	1.000	0	0	0	1.000	0	0	0	1.000	0	0	0
SMB⋆	0	0	0	0.011	0	0	0	0.007	0	0	0	0.006
HML⋆	0	0	0	0.020	0	0	0	0.020	0	0	0	0.020
RMW⋆	0	0	0	0.003	0	0	0	0.002	0	0	0	0.002
CMA⋆	0	0	0	0.002	0	0	0	0.002	0	0	0	0.002
MOM	0	0	0	0.002	0	0	0	0.001	0	0	0	0.001
ME	0	0	0	0.001	0	0	0	0	0	0	0	0
IA	0	0	0	0.001	0	0	0	0	0	0	0	0
ROE	0	0	0	0.001	0	0	0	0.001	0	0	0	0
EG	0	0	0	0	0	0	0	0	0	0	0	0
MGMT	0	0	0	0	0	0	0	0	0	0	0	0
PERF	0	0	0	0.006	0	0	0	0.004	0	0	0	0.004
PEAD	0	0	0	0.007	0	0	0	0.005	0	0	0	0.004
FIN	0	0	0	0	0	0	0	0	0	0	0	0
Ident. prob.	1.000	0.928	0.927	0.705	1.000	0.928	0.951	0.691	1.000	0.924	0.958	0.677

This table presents the Bayesian estimates of the statistical and characteristic-based factors. The data are at the monthly frequency, spanning from January 1974 to December 2019. In empirical applications, all variables are standardized to have unit variance per period. The data description is in Section 3.1. In Panel A, we use the Fama-French five factors (Fama and French, 2015), whereas the hedged returns (Daniel et al., 2020) of the Fama-French five factors are employed in Panel B. We use Algorithm 1, with $L = 10$ and $\sigma_0 \in \{0.3/\sqrt{12}, 0.4/\sqrt{12}, 0.5/\sqrt{12}\}$, to estimate the posterior factor probabilities within identities ($\{\boldsymbol{q}_\ell\}_{\ell=1}^L$). The last row reports the posterior identity probabilities ($\{\theta_\ell\}_{\ell=1}^L$). We display the identities with posterior probabilities above 50%.

Table A8: Statistical vs. Macro Factors, Alternative Priors

	Factor probs. within identities, $\boldsymbol{q}_\ell$						$p(\boldsymbol{m}_\ell \mid \text{data})$	$\mathbb{E}[\boldsymbol{\lambda} \mid \text{data}]$
	$\ell=1$	$\ell=2$	$\ell=3$	$\ell=4$	$\ell=5$	$\ell=6$	mean & 90% CI	
	Panel A. $\sigma_0 = 0.3/\sqrt{4}$							
$\mathbf{1}_n$	0.896	0	0	0	0	0	0.181 (0.047, 0.318)	0.162
PC1	0.044	0	0	0	0	0	0.199 (0.051, 0.353)	0.009
PC2	0	0	0	0	0	0		0
PC3	0	0	0	0	0	0		0.001
PC4	0	0	0	0	0.101	0	0.131 (-0.001, 0.264)	0.014
PC5	0	0	0	0	0	0		-0.006
RP-PC1	0.056	0	0	0	0	0	0.199 (0.051, 0.353)	0.011
RP-PC2	0	1.000	0	0	0	0	0.335 (0.190, 0.485)	0.337
RP-PC3	0	0	0	1.000	0	0	0.167 (0.020, 0.314)	0.168
RP-PC4	0	0	0	0	0.899	0	0.144 (0.003, 0.283)	0.131
RP-PC5	0	0	1.000	0	0	0	0.225 (0.084, 0.369)	0.229
Nondurable (PJ12)	0	0	0	0	0	0.001		0.003
GDP (PJ12)	0	0	0	0	0	0.159	0.175 (0.053, 0.311)	0.033
Industrial production (PJ12)	0	0	0	0	0	0.064	0.169 (0.036, 0.339)	0.016
Nondurable+service (PJ12)	0	0	0	0	0	0.759	0.227 (0.084, 0.372)	0.186
Durable (PJ12)	0	0	0	0	0	0		0.003
Investment (PJ12)	0	0	0	0	0	0.008		0.004
Hours worked (PJ12)	0	0	0	0	0	0.007		0.003
Unemp. rate (PJ12)	0	0	0	0	0	0.002		-0.001
TFP (PJ12)	0	0	0	0	0	0		0.048
Liquidity-ntr	0	0	0	0	0	0		0
HKM-ntr	0.004	0	0	0	0	0		0.001
Ident. prob.	0.990	0.999	0.897	0.866	0.744	0.860		
$\mathbb{E}[\lambda_\ell \mid \text{data}]$	0.183	0.335	0.225	0.167	0.143	0.214		
	Panel B. $\sigma_0 = 0.5/\sqrt{4}$							
$\mathbf{1}_n$	0.752	0	0	0	0	0	0.165 (0.026, 0.311)	0.124
PC1	0.103	0	0	0	0	0	0.183 (0.029, 0.343)	0.019
PC2	0	0	0	0	0	0		0
PC3	0	0	0	0	0	0		0
PC4	0	0	0	0	0.037	0	0.160 (0.009, 0.309)	0.006
PC5	0	0	0	0	0	0		-0.026
RP-PC1	0.116	0	0	0	0	0	0.183 (0.029, 0.343)	0.021
RP-PC2	0	1.000	0	0	0	0	0.362 (0.208, 0.522)	0.362
RP-PC3	0	0	0	1.000	0	0	0.190 (0.032, 0.345)	0.191
RP-PC4	0	0	0	0	0.963	0	0.176 (0.014, 0.334)	0.170
RP-PC5	0	0	1.000	0	0	0	0.272 (0.113, 0.434)	0.273
Nondurable (PJ12)	0	0	0	0	0	0		0.002
GDP (PJ12)	0	0	0	0	0	0.389	0.290 (0.085, 0.511)	0.116
Industrial production (PJ12)	0	0	0	0	0	0.588	0.289 (0.078, 0.516)	0.173
Nondurable+service (PJ12)	0	0	0	0	0	0.004		0.011
Durable (PJ12)	0	0	0	0	0	0		0.001
Investment (PJ12)	0	0	0	0	0	0.013	0.253 (0.057, 0.497)	0.005
Hours worked (PJ12)	0	0	0	0	0	0.006		0.003
Unemp. rate (PJ12)	0	0	0	0	0	0		-0.001
TFP (PJ12)	0	0	0	0	0	0		0.051
Liquidity-ntr	0.002	0	0	0	0	0		0.001
HKM-ntr	0.028	0	0	0	0	0	0.231 (0.036, 0.431)	0.006
Ident. prob.	0.986	0.999	0.928	0.884	0.784	0.920		
$\mathbb{E}[\lambda_\ell \mid \text{data}]$	0.171	0.362	0.272	0.190	0.175	0.289		

This table presents the Bayesian estimates of the statistical and macro factors. The data are at the quarterly frequency, spanning from 1974Q1 to 2019Q4. In empirical applications, all variables are standardized to have unit variance per period. The data description is in Section 3.1. Macro factors with the notation "(PJ12)" are defined as the 12-quarter cumulative growth rates of the underlying economic variables between quarter $t-1$ and quarter $t+12$, following Parker and Julliard (2005). We use Algorithm 1, with $L = 10$ and $\sigma_0 \in \{0.3/\sqrt{4}, 0.5/\sqrt{4}\}$, to estimate the posterior factor probabilities within identities ($\{\boldsymbol{q}_\ell\}_{\ell=1}^L$). The second last row reports the posterior identity probabilities ($\{\theta_\ell\}_{\ell=1}^L$). We display the identities with posterior probabilities above 50%. Conditional on $\{\boldsymbol{q}_\ell\}_{\ell=1}^L$, we estimate the posterior distribution of factor risk prices using the approach described in Section 1.3.2. We present both conditional and unconditional factor risk prices. In the column "$p(\boldsymbol{m}_\ell \mid \text{data})$", we report the factor risk prices conditional on the factors being selected (both the conditional posterior mean and 90% credible intervals). In the column "$\mathbb{E}[\boldsymbol{\lambda} \mid \text{data}]$", we display the unconditional posterior mean of the factor risk prices, defined as $\mathbb{E}[\sum_{\ell=1}^L \boldsymbol{m}_\ell \odot \boldsymbol{q}_\ell \mid \text{data}]$. In the row labeled "$\mathbb{E}[\lambda_\ell \mid \text{data}]$", we report the posterior mean of the identity market price of risk, defined as $\mathbb{E}[\sum_j q_{\ell,j} m_{\ell,j} \mid \text{data}]$.

Table A9: Statistical vs. Macro Factors, Quarterly Growth Rates of Macro Variables

	Panel A. $\sigma_0 = 0.3/\sqrt{4}$				Panel B. $\sigma_0 = 0.4/\sqrt{4}$					Panel C. $\sigma_0 = 0.5/\sqrt{4}$				
	$\ell=1$	$\ell=2$	$\ell=3$	$\ell=4$	$\ell=1$	$\ell=2$	$\ell=3$	$\ell=4$	$\ell=5$	$\ell=1$	$\ell=2$	$\ell=3$	$\ell=4$	$\ell=5$
$\mathbf{1}_n$	0.793	0	0	0	0.779	0	0	0	0	0.728	0	0	0	0
PC1	0.087	0	0	0	0.088	0	0	0	0	0.107	0	0	0	0
PC2	0	0	0.002	0	0	0	0.001	0	0	0	0	0.001	0	0
PC3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PC4	0	0	0	0	0	0	0	0	0.143	0	0	0	0	0.106
PC5	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RP-PC1	0.114	0	0	0	0.113	0	0	0	0	0.135	0	0	0	0
RP-PC2	0	1.000	0	0	0	1.000	0	0	0	0	1.000	0	0	0
RP-PC3	0	0	0.998	0	0	0	0.999	0	0	0	0	0.999	0	0
RP-PC4	0	0	0	0	0	0	0	0	0.856	0	0	0	0	0.894
RP-PC5	0	0	0	1.000	0	0	0	1.000	0	0	0	0	1.000	0
Nondurable	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GDP	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Industrial production	0	0	0	0	0	0	0	0	0.001	0	0	0	0	0
Nondurable+service	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Durable	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Investment	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Hours worked	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Unemp. rate	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TFP	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Liquidity-ntr	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HKM-ntr	0.006	0	0	0	0.021	0	0	0	0	0.030	0	0	0	0
Ident. prob.	0.997	0.998	0.837	0.908	0.997	0.999	0.851	0.958	0.720	0.996	0.999	0.853	0.976	0.740

This table presents the Bayesian estimates of the statistical and macro factors. The data are at the quarterly frequency, spanning from 1974Q1 to 2019Q4. In empirical applications, all variables are standardized to have unit variance per period. The data description is in Section 3.1. We use Algorithm 1, with $L = 10$ and $\sigma_0 \in \{0.3/\sqrt{4}, 0.4/\sqrt{4}, 0.5/\sqrt{4}\}$, to estimate the posterior factor probabilities within identities ($\{\boldsymbol{q}_\ell\}_{\ell=1}^L$). The last row reports the posterior identity probabilities ($\{\theta_\ell\}_{\ell=1}^L$). We display the identities with posterior probabilities above 50%.

Table A10: Characteristic-Based vs. Macro Factors, Alternative Priors

	Panel A. $\sigma_0 = 0.3/\sqrt{4}$						Panel B. $\sigma_0 = 0.5/\sqrt{4}$					
	Factor probs. within identities, $\boldsymbol{q}_\ell$				$p(\boldsymbol{m}_\ell \mid \text{data})$	$\mathbb{E}[\boldsymbol{\lambda} \mid \text{data}]$	Factor probs. within identities, $\boldsymbol{q}_\ell$				$p(\boldsymbol{m}_\ell \mid \text{data})$	$\mathbb{E}[\boldsymbol{\lambda} \mid \text{data}]$
	$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 4$	mean & with 90% CI		$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 4$	mean & 90% CI	
$\mathbf{1}_n$	0	0	0	0		0	0	0	0	0		0
MKT	0	0	0	0		0	0	0	0	0		0
SMB	0	0	0	0		0	0	0	0	0		0
HML	0	0	0	0		0.001	0	0	0	0		0.001
RMW	0	0	0	0		0	0	0	0	0		0
CMA	0	0	0	0		0	0	0	0	0		0.001
MOM	0	0	0.457	0	0.173 (0.016, 0.330)	0.080	0	0	0.076	0	0.248 (0.057, 0.465)	0.021
MKT$\star$	1.000	0	0	0	0.380 (0.150, 0.610)	0.380	1.000	0	0	0	0.396 (0.148, 0.643)	0.396
SMB$\star$	0	1.000	0	0	0.306 (0.125, 0.484)	0.309	0	1.000	0	0	0.361 (0.140, 0.581)	0.363
HML$\star$	0	0	0	0		0.001	0	0	0	0		0.002
RMW$\star$	0	0	0	0		-0.001	0	0	0	0		-0.003
CMA$\star$	0	0	0	0		0.005	0	0	0	0		0.014
ME	0	0	0	0		0.001	0	0	0	0		0
IA	0	0	0	0		0	0	0	0	0		0.001
ROE	0	0	0	0		0	0	0	0	0		0
EG	0	0	0	0		0	0	0	0	0		0
MGMT	0	0	0	0		0	0	0	0	0		0
PERF	0	0	0.542	0	0.172 (0.018, 0.332)	0.094	0	0	0.924	0	0.249 (0.051, 0.444)	0.231
PEAD	0	0	0	0		0.003	0	0	0	0		0.010
FIN	0	0	0	0		0	0	0	0	0		0
Nondurable (PJ12)	0	0	0	0		0.003	0	0	0	0		-0.002
GDP (PJ12)	0	0	0	0.096	0.214 (0.054, 0.398)	0.027	0	0	0	0.027	0.310 (0.065, 0.600)	0.013
Industrial production (PJ12)	0	0	0	0.027	0.207 (0.038, 0.428)	0.011	0	0	0	0.004		0.006
Nondurable+service (PJ12)	0	0	0	0.874	0.281 (0.076, 0.475)	0.264	0	0	0	0.969	0.420 (0.115, 0.713)	0.419
Durable (PJ12)	0	0	0	0		0.005	0	0	0	0		0.002
Investment (PJ12)	0	0	0	0.001		0.002	0	0	0	0		0.001
Hours worked (PJ12)	0	0	0	0.001		0.002	0	0	0	0		0
Unemp. rate (PJ12)	0	0	0	0		-0.001	0	0	0	0		0.001
TFP (PJ12)	0	0	0	0		0.145	0	0	0	0		0.135
Liquidity-ntr	0	0	0	0		0	0	0	0	0		0
HKM-ntr	0	0	0	0		0	0	0	0	0		0
Ident. prob.	0.998	0.966	0.916	0.850			0.997	0.967	0.956	0.892		
$\mathbb{E}[\lambda_\ell \mid \text{data}]$	0.380	0.306	0.173	0.272			0.396	0.361	0.249	0.417		

This table presents the Bayesian estimates of the characteristic-based and macro factors. The data are at the quarterly frequency, spanning from 1974Q1 to 2019Q4. In empirical applications, all variables are standardized to have unit variance per period. The data description is in Section 3.1. Macro factors with the notation "(PJ12)" are defined as the 12-quarter cumulative growth rates of the underlying economic variables between quarter $t-1$ and quarter $t+12$, following Parker and Julliard (2005). We use Algorithm 1, with $L = 10$ and $\sigma_0 \in \{0.3/\sqrt{4}, 0.5/\sqrt{4}\}$, to estimate the posterior factor probabilities within identities ($\{\boldsymbol{q}_\ell\}_{\ell=1}^L$). The second last row reports the posterior identity probabilities ($\{\theta_\ell\}_{\ell=1}^L$). We display the identities with posterior probabilities above 50%. Conditional on $\{\boldsymbol{q}_\ell\}_{\ell=1}^L$, we estimate the posterior distribution of factor risk prices using the approach described in Section 1.3.2. We present both conditional and unconditional factor risk prices. In the column "$p(\boldsymbol{m}_\ell \mid \text{data})$", we report the factor risk prices conditional on the factors being selected (both the conditional posterior mean and 90% credible intervals). In the column "$\mathbb{E}[\boldsymbol{\lambda} \mid \text{data}]$", we display the unconditional posterior mean of the factor risk prices, defined as $\mathbb{E}[\sum_{\ell=1}^L \boldsymbol{m}_\ell \odot \boldsymbol{q}_\ell \mid \text{data}]$. In the row labeled "$\mathbb{E}[\lambda_\ell \mid \text{data}]$", we report the posterior mean of the identity market price of risk, defined as $\mathbb{E}[\sum_j q_{\ell,j} m_{\ell,j} \mid \text{data}]$.

Table A11: Statistical vs. Characteristic-Based vs. Macro Factors, Alternative Priors

	Panel A. $\sigma_0 = 0.3/\sqrt{4}$							Panel B. $\sigma_0 = 0.5/\sqrt{4}$						
	Factor probs. within identities, $\boldsymbol{q}_\ell$					$p(\boldsymbol{m}_\ell \mid \text{data})$	$\mathbb{E}[\boldsymbol{\lambda} \mid \text{data}]$	Factor probs. within identities, $\boldsymbol{q}_\ell$					$p(\boldsymbol{m}_\ell \mid \text{data})$	$\mathbb{E}[\boldsymbol{\lambda} \mid \text{data}]$
	$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 4$	$\ell = 5$	mean & 90% CI		$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 4$	$\ell = 5$	mean & 90% CI	
$\mathbf{1}_n$	0	0	0	0	0		0	0.949	0	0	0	0	0.169 (0.019, 0.313)	0.160
PC1	0	0	0	0	0		0	0.015	0	0	0	0	0.186 (0.020, 0.346)	0.003
PC2	0	0	0	0	0		0	0	0	0	0	0		0.005
PC3	0	0	0	0	0		0.001	0	0	1.000	0	0	0.375 (0.206, 0.547)	0.376
PC4	0	0	0.002	0	0		0.002	0	0	0	0.014	0	0.215 (0.064, 0.378)	0.003
PC5	0	0	0	0	0		-0.001	0	0	0	0	0		-0.012
RP-PC1	0	0	0	0	0		0	0.019	0	0	0	0	0.186 (0.020, 0.346)	0.004
RP-PC2	0	1.000	0	0	0	0.246 (0.068, 0.420)	0.247	0	0	0	0	0		0.007
RP-PC3	0	0	0	0	0		0.001	0	0	0	0	0		-0.003
RP-PC4	0	0	0.998	0	0	0.274 (0.105, 0.459)	0.276	0	0	0	0.986	0	0.235 (0.073, 0.406)	0.233
RP-PC5	0	0	0	1.000	0	0.251 (0.103, 0.406)	0.256	0	1.000	0	0	0	0.303 (0.138, 0.474)	0.304
MKT	0	0	0	0	0		0	0.007	0	0	0	0		0.001
SMB	0	0	0	0	0		0	0	0	0	0	0		0
HML	0	0	0	0	0		0	0	0	0	0	0		0.001
RMW	0	0	0	0	0		0	0	0	0	0	0		0
CMA	0	0	0	0	0		0	0	0	0	0	0		0
MOM	0	0	0	0	0		0	0	0	0	0	0		-0.001
MKT$\star$	1.000	0	0	0	0	0.387 (0.134, 0.644)	0.387	0	0	0	0	0		0
SMB$\star$	0	0	0	0	0		0.001	0	0	0	0	0		0.002
HML$\star$	0	0	0	0	0		0	0	0	0	0	0		0.002
RMW$\star$	0	0	0	0	0		-0.001	0	0	0	0	0		-0.002
CMA$\star$	0	0	0	0	0		0	0	0	0	0	0		0.007
ME	0	0	0	0	0		0	0	0	0	0	0		0
IA	0	0	0	0	0		0	0	0	0	0	0		0
ROE	0	0	0	0	0		0	0	0	0	0	0		0
EG	0	0	0	0	0		0	0	0	0	0	0		0
MGMT	0	0	0	0	0		0	0	0	0	0	0		0
PERF	0	0	0	0	0		0	0	0	0	0	0		-0.001
PEAD	0	0	0	0	0		0	0	0	0	0	0		-0.003
FIN	0	0	0	0	0		0	0	0	0	0	0		0
Nondurable (PJ12)	0	0	0	0	0.001		0.003	0	0	0	0	0		0
GDP (PJ12)	0	0	0	0	0.120	0.177 (0.017, 0.360)	0.026	0	0	0	0	0.070	0.283 (0.064, 0.540)	0.023
Industrial production (PJ12)	0	0	0	0	0.059	0.170 (0.003, 0.397)	0.015	0	0	0	0	0.010		0.007
Nondurable+service (PJ12)	0	0	0	0	0.809	0.226 (0.022, 0.439)	0.193	0	0	0	0	0.920	0.374 (0.117, 0.632)	0.351
Durable (PJ12)	0	0	0	0	0.001		0.003	0	0	0	0	0		0
Investment (PJ12)	0	0	0	0	0.005		0.003	0	0	0	0	0		0.002
Hours worked (PJ12)	0	0	0	0	0.004		0.003	0	0	0	0	0		0.001
Unemp. rate (PJ12)	0	0	0	0	0.001		-0.002	0	0	0	0	0		-0.001
TFP (PJ12)	0	0	0	0	0		0.078	0	0	0	0	0		0.015
Liquidity-ntr	0	0	0	0	0		0	0	0	0	0	0		0
HKM-ntr	0	0	0	0	0		0	0.009	0	0	0	0		0.002
Ident. prob.	0.996	0.952	0.952	0.916	0.804			0.984	0.949	0.996	0.901	0.914		
$\mathbb{E}[\lambda_\ell \mid \text{data}]$	0.387	0.246	0.274	0.251	0.216			0.170	0.303	0.375	0.235	0.367		

This table presents the Bayesian estimates of the statistical, characteristic-based, and macro factors. The data are at the quarterly frequency, spanning from 1974Q1 to 2019Q4. In empirical applications, all variables are standardized to have unit variance per period. The data description is in Section 3.1. Macro factors with the notation "(PJ12)" are defined as the 12-quarter cumulative growth rates of the underlying economic variables between quarter $t-1$ and quarter $t+12$, following Parker and Julliard (2005). We use Algorithm 1, with $L = 10$ and $\sigma_0 \in \{0.3/\sqrt{4}, 0.5/\sqrt{4}\}$, to estimate the posterior factor probabilities within identities ($\{\boldsymbol{q}_\ell\}_{\ell=1}^L$). The second last row reports the posterior identity probabilities ($\{\theta_\ell\}_{\ell=1}^L$). We display the identities with posterior probabilities above 50%. Conditional on $\{\boldsymbol{q}_\ell\}_{\ell=1}^L$, we estimate the posterior distribution of factor risk prices using the approach described in Section 1.3.2. We present both conditional and unconditional factor risk prices. In the column "$p(\boldsymbol{m}_\ell \mid \text{data})$", we report the factor risk prices conditional on the factors being selected (both the conditional posterior mean and 90% credible intervals). In the column "$\mathbb{E}[\boldsymbol{\lambda} \mid \text{data}]$", we display the unconditional posterior mean of the factor risk prices, defined as $\mathbb{E}[\sum_{\ell=1}^L \boldsymbol{m}_\ell \odot \boldsymbol{q}_\ell \mid \text{data}]$. In the row labeled "$\mathbb{E}[\lambda_\ell \mid \text{data}]$", we report the posterior mean of the identity market price of risk, defined as $\mathbb{E}[\sum_j q_{\ell,j} m_{\ell,j} \mid \text{data}]$.

Table A12: Statistical Factors vs. Value-Related Anomalies

	Panel A. $\sigma_0 = 0.3/\sqrt{12}$				Panel B. $\sigma_0 = 0.4/\sqrt{12}$				Panel C. $\sigma_0 = 0.5/\sqrt{12}$			
	$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 4$	$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 4$	$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 4$
$\mathbf{1}_n$	0	0	0	0	0	0	0	0	0	0	0	0
PC1	0	0	0	0	0	0	0	0	0	0	0	0
PC2	0	0	0	0.110	0	0	0	0.061	0	0	0	0.028
PC3	0	0	0	0	0	0	0	0	0	0	0	0
PC4	0	0	0.995	0	0	0	0.998	0	0	0	0.999	0
PC5	0	0	0	0	0	0	0	0	0	0	0	0
RP-PC1	0	0	0	0	0	0	0	0	0	0	0	0
RP-PC2	0	0	0	0.002	0	0	0	0	0	0	0	0
RP-PC3	0	0	0	0.002	0	0	0	0.001	0	0	0	0
RP-PC4	0	0	0.005	0	0	0	0.002	0	0	0	0.001	0
RP-PC5	0	1.000	0	0	0	1.000	0	0	0	1.000	0	0
MKT	0	0	0	0	0	0	0	0	0	0	0	0
MKT*	1.000	0	0	0	1.000	0	0	0	1.000	0	0	0
value	0	0	0	0.490	0	0	0	0.518	0	0	0	0.470
valuem	0	0	0	0	0	0	0	0	0	0	0	0
divp	0	0	0	0	0	0	0	0	0	0	0	0
ep	0	0	0	0	0	0	0	0	0	0	0	0
cfp	0	0	0	0.127	0	0	0	0.181	0	0	0	0.314
dur	0	0	0	0.258	0	0	0	0.236	0	0	0	0.187
sp	0	0	0	0.005	0	0	0	0.001	0	0	0	0
valmom	0	0	0	0	0	0	0	0	0	0	0	0
valmomprof	0	0	0	0	0	0	0	0	0	0	0	0
valprof	0	0	0	0.004	0	0	0	0.001	0	0	0	0.001
Ident. prob.	1.000	0.998	0.983	0.833	1.000	0.999	0.985	0.860	1.000	1.000	0.986	0.881

This table presents the Bayesian estimates for the statistical, market, and value-related anomaly factors. The data are monthly and span the period from January 1974 to December 2019. In empirical applications, all variables are standardized to have unit variance per period. The data description is in Section 3.1. We use Algorithm 1, with $L = 10$ and $\sigma_0 \in \{0.3/\sqrt{12}, 0.4/\sqrt{12}, 0.5/\sqrt{12}\}$, to estimate the posterior factor probabilities within identities ($\{\boldsymbol{q}_\ell\}_{\ell=1}^L$). The last row reports the posterior identity probabilities ($\{\theta_\ell\}_{\ell=1}^L$). We display the identities with posterior probabilities above 50%.

Table A13: Statistical Factors vs. Anomalies based on Technical Signals

	Panel A. $\sigma_0 = 0.3/\sqrt{12}$				Panel B. $\sigma_0 = 0.4/\sqrt{12}$					Panel C. $\sigma_0 = 0.5/\sqrt{12}$					
	$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 4$	$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 4$	$\ell = 5$	$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 4$	$\ell = 5$	$\ell = 6$
$\mathbf{1}_n$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PC1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PC2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.036
PC3	0	0	0	0	0	0	0	0	0	0	0.001	0	0	0	0
PC4	0	0.002	0	0	0	0.042	0	0	0	0	0.262	0	0	0	0
PC5	0	0	0	0	0	0	0	0	0	0	0.001	0	0	0	0
RP-PC1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RP-PC2	0	0	0	1.000	0	0	0	1.000	0	0	0	0	1.000	0	0
RP-PC3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.089
RP-PC4	0	0.998	0	0	0	0.957	0	0	0	0	0.694	0	0	0	0
RP-PC5	0	0	1.000	0	0	0	1.000	0	0	0	0.001	1.000	0	0	0
MKT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MKT*	1.000	0	0	0	1.000	0	0	0	0	1.000	0	0	0	0	0
lrrev	0	0	0	0	0	0	0	0	0	0	0.001	0	0	0	0.676
strev	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
indrrev	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
indmomrev	0	0	0	0	0	0	0	0	1.000	0	0.033	0	0	1.000	0
indrrevlv	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.001
mom	0	0	0	0	0	0	0	0	0	0	0.001	0	0	0	0
mom12	0	0	0	0	0	0	0	0	0	0	0.001	0	0	0	0
indmom	0	0	0	0	0	0	0	0	0	0	0.001	0	0	0	0
momrev	0	0	0	0	0	0	0	0	0	0	0.001	0	0	0	0.197
Ident. prob.	0.999	0.979	0.999	0.930	0.999	0.898	1.000	0.906	0.863	0.998	0.793	1.000	0.830	0.953	0.742

This table presents the Bayesian estimates for the statistical, market, and momentum-and-reversal anomaly factors. The data are monthly and span the period from January 1974 to December 2019. In empirical applications, all variables are standardized to have unit variance per period. The data description is in Section 3.1. We use Algorithm 1, with $L = 10$ and $\sigma_0 \in \{0.3/\sqrt{12}, 0.4/\sqrt{12}, 0.5/\sqrt{12}\}$, to estimate the posterior factor probabilities within identities ($\{\boldsymbol{q}_\ell\}_{\ell=1}^L$). The last row reports the posterior identity probabilities ($\{\theta_\ell\}_{\ell=1}^L$). We display the identities with posterior probabilities above 50%.

Table A14: Statistical Factors vs. Anomalies in Other Categories

	Panel A. $\sigma_0 = 0.3/\sqrt{12}$				Panel B. $\sigma_0 = 0.4/\sqrt{12}$				Panel C. $\sigma_0 = 0.5/\sqrt{12}$			
	$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 4$	$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 4$	$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 4$
Investment Anomalies												
$\mathbf{1}_n$	0	0	0	0	0	0	0	0	0	0	0	0
PC1	0	0	0	0	0	0	0	0	0	0	0	0
PC2	0	0	0	0	0	0	0	0	0	0	0	0
PC3	0	0	0	0	0	0	0	0	0	0	0	0
PC4	0	0.001	0	0	0	0	0	0	0	0	0	0
PC5	0	0	0	0	0	0	0	0	0	0	0	0
RP-PC1	0	0	0	0	0	0	0	0	0	0	0	0
RP-PC2	0	0	0	1.000	0	0	0	1.000	0	0	0	1.000
RP-PC3	0	0	0	0	0	0	0	0	0	0	0	0
RP-PC4	0	0.999	0	0	0	1.000	0	0	0	1.000	0	0
RP-PC5	0	0	1.000	0	0	0	1.000	0	0	0	1.000	0
MKT	0	0	0	0	0	0	0	0	0	0	0	0
MKT*	1.000	0	0	0	1.000	0	0	0	1.000	0	0	0
inv	0	0	0	0	0	0	0	0	0	0	0	0
invcap	0	0	0	0	0	0	0	0	0	0	0	0
igrowth	0	0	0	0	0	0	0	0	0	0	0	0
growth	0	0	0	0	0	0	0	0	0	0	0	0
noa	0	0	0	0	0	0	0	0	0	0	0	0
gltnoa	0	0	0	0	0	0	0	0	0	0	0	0
sgrowth	0	0	0	0	0	0	0	0	0	0	0	0
Ident. prob.	0.999	0.979	0.999	0.921	0.999	0.981	1.000	0.921	0.999	0.982	1.000	0.921
Profitability Anomalies												
$\mathbf{1}_n$	0	0	0	0	0	0	0	0	0	0	0	0
PC1	0	0	0	0	0	0	0	0	0	0	0	0
PC2	0	0	0	0	0	0	0	0	0	0	0	0
PC3	0	0	0	0	0	0	0	0	0	0	0	0
PC4	0	0.001	0	0	0	0	0	0	0	0	0	0
PC5	0	0	0	0	0	0	0	0	0	0	0	0
RP-PC1	0	0	0	0	0	0	0	0	0	0	0	0
RP-PC2	0	0	0	1.000	0	0	0	1.000	0	0	0	1.000
RP-PC3	0	0	0	0	0	0	0	0	0	0	0	0
RP-PC4	0	0.999	0	0	0	1.000	0	0	0	1.000	0	0
RP-PC5	0	0	1.000	0	0	0	1.000	0	0	0	1.000	0
MKT	0	0	0	0	0	0	0	0	0	0	0	0
MKT*	1.000	0	0	0	1.000	0	0	0	1.000	0	0	0
prof	0	0	0	0	0	0	0	0	0	0	0	0
roaa	0	0	0	0	0	0	0	0	0	0	0	0
roea	0	0	0	0	0	0	0	0	0	0	0	0
roa	0	0	0	0	0	0	0	0	0	0	0	0
roe	0	0	0	0	0	0	0	0	0	0	0	0
rome	0	0	0	0	0	0	0	0	0	0	0	0
gmargins	0	0	0	0	0	0	0	0	0	0	0	0
aturnover	0	0	0	0	0	0	0	0	0	0	0	0
Ident. prob.	0.999	0.980	1.000	0.923	0.999	0.984	1.000	0.930	0.999	0.987	1.000	0.936
Trading-Friction Anomalies												
$\mathbf{1}_n$	0	0	0	0	0	0	0	0	0	0	0	0
PC1	0	0	0	0	0	0	0	0	0	0	0	0
PC2	0	0	0	0	0	0	0	0	0	0	0	0
PC3	0	0	0	0	0	0	0	0	0	0	0	0
PC4	0	0.001	0	0	0	0	0	0	0	0	0	0
PC5	0	0	0	0	0	0	0	0	0	0	0	0
RP-PC1	0	0	0	0	0	0	0	0	0	0	0	0
RP-PC2	0	0	0	1.000	0	0	0	1.000	0	0	0	1.000
RP-PC3	0	0	0	0	0	0	0	0	0	0	0	0
RP-PC4	0	0.999	0	0	0	1.000	0	0	0	1.000	0	0

Continued on next page

Table A14 – continued from previous page

	Panel A. $\sigma_0 = 0.3/\sqrt{12}$				Panel B. $\sigma_0 = 0.4/\sqrt{12}$				Panel C. $\sigma_0 = 0.5/\sqrt{12}$			
	$\ell=1$	$\ell=2$	$\ell=3$	$\ell=4$	$\ell=1$	$\ell=2$	$\ell=3$	$\ell=4$	$\ell=1$	$\ell=2$	$\ell=3$	$\ell=4$
RP-PC5	0	0	1.000	0	0	0	1.000	0	0	0	1.000	0
MKT	0	0	0	0	0	0	0	0	0	0	0	0
MKT*	1.000	0	0	0	1.000	0	0	0	1.000	0	0	0
ivol	0	0	0	0	0	0	0	0	0	0	0	0
shvol	0	0	0	0	0	0	0	0	0	0	0	0
betaarb	0	0	0	0	0	0	0	0	0	0	0	0
sue	0	0	0	0	0	0	0	0	0	0	0	0
shortint	0	0	0	0	0	0	0	0	0	0	0	0
Ident. prob.	0.999	0.978	0.999	0.926	0.999	0.981	1.000	0.926	0.999	0.983	1.000	0.927
Issuance Anomalies												
$\mathbf{1}_n$	0	0	0	0	0	0	0	0	0	0	0	0
PC1	0	0	0	0	0	0	0	0	0	0	0	0
PC2	0	0	0	0	0	0	0	0	0	0	0	0
PC3	0	0	0	0	0	0	0	0	0	0	0	0
PC4	0	0.001	0	0	0	0	0	0	0	0	0	0
PC5	0	0	0	0	0	0	0	0	0	0	0	0
RP-PC1	0	0	0	0	0	0	0	0	0	0	0	0
RP-PC2	0	0	0	1.000	0	0	0	1.000	0	0	0	1.000
RP-PC3	0	0	0	0	0	0	0	0	0	0	0	0
RP-PC4	0	0.999	0	0	0	1.000	0	0	0	1.000	0	0
RP-PC5	0	0	1.000	0	0	0	1.000	0	0	0	1.000	0
MKT	0	0	0	0	0	0	0	0	0	0	0	0
MKT*	1.000	0	0	0	1.000	0	0	0	1.000	0	0	0
repurch	0	0	0	0	0	0	0	0	0	0	0	0
debtiss	0	0	0	0	0	0	0	0	0	0	0	0
ciss	0	0	0	0	0	0	0	0	0	0	0	0
nissa	0	0	0	0	0	0	0	0	0	0	0	0
nissm	0	0	0	0	0	0	0	0	0	0	0	0
Ident. prob.	0.999	0.980	0.999	0.929	0.999	0.983	1.000	0.933	0.999	0.984	1.000	0.936
Other Anomalies												
$\mathbf{1}_n$	0	0	0	0	0	0	0	0	0	0	0	0
PC1	0	0	0	0	0	0	0	0	0	0	0	0
PC2	0	0	0	0	0	0	0	0	0	0	0	0
PC3	0	0	0	0	0	0	0	0	0	0	0	0
PC4	0	0.001	0	0	0	0	0	0	0	0	0	0
PC5	0	0	0	0	0	0	0	0	0	0	0	0
RP-PC1	0	0	0	0	0	0	0	0	0	0	0	0
RP-PC2	0	0	0	1.000	0	0	0	1.000	0	0	0	1.000
RP-PC3	0	0	0	0	0	0	0	0	0	0	0	0
RP-PC4	0	0.999	0	0	0	1.000	0	0	0	1.000	0	0
RP-PC5	0	0	1.000	0	0	0	1.000	0	0	0	1.000	0
MKT	0	0	0	0	0	0	0	0	0	0	0	0
MKT*	1.000	0	0	0	1.000	0	0	0	1.000	0	0	0
size	0	0	0	0	0	0	0	0	0	0	0	0
price	0	0	0	0	0	0	0	0	0	0	0	0
accruals	0	0	0	0	0	0	0	0	0	0	0	0
age	0	0	0	0	0	0	0	0	0	0	0	0
fscore	0	0	0	0	0	0	0	0	0	0	0	0
lev	0	0	0	0	0	0	0	0	0	0	0	0
season	0	0	0	0	0	0	0	0	0	0	0	0
Ident. prob.	0.999	0.982	1.000	0.930	0.999	0.984	1.000	0.929	0.999	0.984	1.000	0.929

This table presents the Bayesian estimates for the statistical, market, and anomaly factors related to investment, profitability, trading frictions, issuance, and other categories. The data are monthly and span the period from January 1974 to December 2019. In empirical applications, all variables are standardized to have unit variance per period. The data description is in Section 3.1. We use Algorithm 1, with $L = 10$ and $\sigma_0 \in \{0.3/\sqrt{12}, 0.4/\sqrt{12}, 0.5/\sqrt{12}\}$, to estimate the posterior factor probabilities within identities ($\{\boldsymbol{q}_\ell\}_{\ell=1}^L$). The last row reports the posterior identity probabilities ($\{\theta_\ell\}_{\ell=1}^L$). We display the identities with posterior probabilities above 50%.

Table A15: Factor Horse Race: 51 Anomalies and Market Factors ($\sigma_0 = 0.3/\sqrt{12}$)

	Factor probs. within identities, $\boldsymbol{q}_\ell$					$p(\boldsymbol{\lambda} \mid \text{data})$
	$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 4$	$\ell = 5$	mean & 90% CI
$\mathbf{1}_n$	0	0	0	0.004	0.004	0.000 (0.000, 0.000)
MKT	0	0	0	0.005	0.004	0.000 (0.000, 0.000)
MKT$\star$	1.000	0	0	0.010	0.009	0.304 (0.162, 0.449)
accruals	0	0	0	0.002	0.042	0.005 (-0.009, 0.018)
aturnover	0	0	0	0.333	0.077	0.054 (0.012, 0.097)
betaarb	0	0	0	0.006	0.006	-0.001 (-0.002, 0.001)
cfp	0	0	0	0.001	0.016	0.001 (-0.001, 0.004)
ciss	0	0	0	0.004	0.007	0.000 (-0.004, 0.005)
nissa	0	0	0	0.002	0.009	0.003 (-0.003, 0.008)
divp	0	0	0	0.012	0.011	-0.001 (-0.002, 0.001)
dur	0	0	0	0.001	0.019	0.001 (-0.001, 0.004)
ep	0	0	0	0.002	0.008	0.001 (-0.003, 0.005)
gmargins	0	0	0	0.002	0.018	-0.003 (-0.010, 0.003)
growth	0	0	0	0.004	0.013	0.003 (-0.001, 0.007)
igrowth	0	0	0	0.005	0.018	0.005 (0.000, 0.012)
indmom	0	0	0	0.004	0.017	0.001 (-0.002, 0.004)
indrrev	0	0	0	0.036	0.017	0.010 (0.003, 0.018)
inv	0	0	0	0.006	0.015	0.006 (0.001, 0.011)
invcap	0	0	0	0.006	0.008	0.000 (-0.004, 0.003)
lev	0	0	0	0.001	0.019	-0.001 (-0.005, 0.003)
lrrev	0	0	0	0.001	0.023	0.002 (-0.001, 0.006)
mom	0	0	0	0.002	0.015	0.001 (-0.002, 0.004)
mom12	0	0	0	0.001	0.013	0.002 (0.000, 0.004)
momrev	0	0	0	0.002	0.032	0.006 (-0.001, 0.015)
noa	0	0	0	0.008	0.044	0.009 (-0.006, 0.025)
price	0	0	0	0.003	0.009	0.000 (-0.002, 0.002)
prof	0	0	0	0.037	0.031	0.006 (-0.005, 0.017)
roaa	0	0	0	0.002	0.008	0.000 (-0.004, 0.004)
roea	0	0	0	0.002	0.008	0.000 (-0.005, 0.004)
season	0	0	0	0.289	0.074	0.051 (0.019, 0.081)
sgrowth	0	0	0	0.006	0.012	0.000 (-0.004, 0.005)
shvol	0	0	0	0.005	0.006	-0.001 (-0.004, 0.002)
size	0	0	0	0.002	0.015	0.001 (-0.002, 0.004)
sp	0	0	0	0.001	0.026	0.003 (-0.001, 0.007)
strev	0	0	0	0.027	0.018	0.006 (0.000, 0.012)
valmom	0	0	0	0.003	0.015	0.002 (-0.001, 0.005)
valmomprof	0	0	0	0.001	0.017	0.004 (0.001, 0.007)
valprof	0	1.000	0	0.001	0.022	0.221 (0.135, 0.307)
value	0	0	0	0.001	0.018	0.001 (-0.001, 0.004)
valuem	0	0	0	0.001	0.019	0.000 (-0.002, 0.002)
indmomrev	0	0	0.999	0.003	0.040	0.189 (0.099, 0.280)
age	0	0	0	0.003	0.007	0.000 (-0.004, 0.005)
debtiss	0	0	0	0.006	0.010	0.001 (-0.007, 0.009)
fscore	0	0	0	0.002	0.013	0.000 (-0.008, 0.007)
gltnoa	0	0	0	0.003	0.050	-0.001 (-0.025, 0.022)
indrrevlv	0	0	0	0.121	0.039	0.045 (0.021, 0.070)
nissm	0	0	0	0.001	0.009	0.002 (-0.004, 0.007)
ivol	0	0	0	0.003	0.006	0.000 (-0.004, 0.003)
repurch	0	0	0	0.002	0.009	0.000 (-0.007, 0.008)
roa	0	0	0	0.001	0.008	0.001 (-0.003, 0.004)
roe	0	0	0	0.001	0.008	0.001 (-0.003, 0.005)
rome	0	0	0	0.001	0.008	0.002 (-0.003, 0.008)
shortint	0	0	0	0.010	0.009	-0.003 (-0.010, 0.003)
sue	0	0	0	0.001	0.024	0.006 (0.000, 0.012)
Ident. prob.	1.000	0.995	0.922	0.699	0.669	

This table presents the Bayesian estimates of the 51 long-short anomalies, with the control for two market factors. The data are at the monthly frequency, spanning from January 1974 to December 2019. In empirical applications, all variables are standardized to have unit variance per period. The data description is in Section 3.1. We use Algorithm 1, with $L = 10$ and $\sigma_0 = 0.3/\sqrt{12}$, to estimate the posterior factor probabilities within identities ($\{\boldsymbol{q}_\ell\}_{\ell=1}^L$). The last row reports the posterior identity probabilities ($\{\theta_\ell\}_{\ell=1}^L$).

Table A16: Factor Horse Race: 51 Anomalies and Market Factors ($\sigma_0 = 0.4/\sqrt{12}$)

	Factor probs. within identities, $\boldsymbol{q}_\ell$					$p(\boldsymbol{\lambda} \mid \text{data})$
	$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 4$	$\ell = 5$	mean & 90% CI
$\mathbf{1}_n$	0	0	0	0	0.004	0.000 (0.000, 0.000)
MKT	0	0	0	0	0.004	0.000 (0.000, 0.000)
MKT$\star$	1.000	0	0	0	0.009	0.311 (0.167, 0.458)
accruals	0	0	0	0	0.048	0.003 (-0.019, 0.023)
aturnover	0	0	0	0	0.070	0.026 (-0.001, 0.054)
betaarb	0	0	0	0	0.006	0.000 (-0.002, 0.001)
cfp	0	0	0	0	0.017	0.001 (-0.002, 0.004)
ciss	0	0	0	0	0.007	0.000 (-0.004, 0.004)
nissa	0	0	0	0	0.010	0.003 (-0.004, 0.010)
divp	0	0	0	0.002	0.011	-0.001 (-0.002, 0.001)
dur	0	0	0	0	0.020	0.001 (-0.001, 0.004)
ep	0	0	0	0	0.009	0.001 (-0.004, 0.005)
gmargins	0	0	0	0	0.018	-0.002 (-0.010, 0.004)
growth	0	0	0	0	0.015	0.002 (-0.002, 0.007)
igrowth	0	0	0	0	0.020	0.006 (-0.002, 0.014)
indmom	0	0	0	0	0.019	0.000 (-0.004, 0.003)
indrrev	0	0	0	0	0.015	0.007 (0.002, 0.013)
inv	0	0	0	0	0.016	0.006 (0.000, 0.012)
invcap	0	0	0	0	0.008	0.000 (-0.004, 0.003)
lev	0	0	0	0	0.021	-0.001 (-0.005, 0.004)
lrrev	0	0	0	0	0.025	0.002 (-0.001, 0.006)
mom	0	0	0	0	0.017	0.000 (-0.003, 0.003)
mom12	0	0	0	0	0.015	0.002 (-0.001, 0.004)
momrev	0	0	0	0	0.034	0.007 (-0.001, 0.016)
noa	0	0	0	0	0.045	0.011 (-0.009, 0.031)
price	0	0	0	0	0.009	0.000 (-0.002, 0.001)
prof	0	0	0	0	0.031	0.005 (-0.005, 0.016)
roaa	0	0	0	0	0.008	0.000 (-0.004, 0.004)
roea	0	0	0	0	0.008	0.000 (-0.005, 0.004)
season	0	0	0	0.997	0.032	0.194 (0.093, 0.291)
sgrowth	0	0	0	0	0.013	0.000 (-0.005, 0.005)
shvol	0	0	0	0	0.006	-0.001 (-0.003, 0.002)
size	0	0	0	0	0.015	0.001 (-0.002, 0.004)
sp	0	0	0	0	0.024	0.002 (-0.001, 0.006)
strev	0	0	0	0	0.016	0.005 (0.000, 0.010)
valmom	0	0	0	0	0.017	0.001 (-0.002, 0.004)
valmomprof	0	0	0	0	0.019	0.003 (0.000, 0.006)
valprof	0	1.000	0	0	0.024	0.306 (0.205, 0.404)
value	0	0	0	0	0.019	0.001 (-0.001, 0.004)
valuem	0	0	0	0	0.019	0.000 (-0.002, 0.002)
indmomrev	0	0	1.000	0	0.036	0.207 (0.096, 0.315)
age	0	0	0	0	0.007	0.000 (-0.004, 0.005)
debtiss	0	0	0	0	0.010	0.002 (-0.008, 0.012)
fscore	0	0	0	0	0.014	0.000 (-0.011, 0.010)
gltnoa	0	0	0	0	0.062	-0.002 (-0.043, 0.039)
indrrevlv	0	0	0	0	0.035	0.031 (0.013, 0.049)
nissm	0	0	0	0	0.011	0.002 (-0.005, 0.009)
ivol	0	0	0	0	0.006	0.000 (-0.003, 0.003)
repurch	0	0	0	0	0.009	0.000 (-0.009, 0.010)
roa	0	0	0	0	0.009	0.001 (-0.004, 0.005)
roe	0	0	0	0	0.009	0.001 (-0.004, 0.005)
rome	0	0	0	0	0.010	0.002 (-0.004, 0.009)
shortint	0	0	0	0	0.008	-0.002 (-0.008, 0.003)
sue	0	0	0	0	0.029	0.007 (-0.002, 0.015)
Ident. prob.	1.000	1.000	0.932	0.874	0.683	

This table presents the Bayesian estimates of the 51 long-short anomalies, with the control for two market factors. The data are at the monthly frequency, spanning from January 1974 to December 2019. In empirical applications, all variables are standardized to have unit variance per period. The data description is in Section 3.1. We use Algorithm 1, with $L = 10$ and $\sigma_0 = 0.4/\sqrt{12}$, to estimate the posterior factor probabilities within identities ($\{\boldsymbol{q}_\ell\}_{\ell=1}^L$). The last row reports the posterior identity probabilities ($\{\theta_\ell\}_{\ell=1}^L$).

Table A17: Factor Horse Race: 51 Anomalies and Market Factors ($\sigma_0 = 0.5/\sqrt{12}$)

	Factor probs. within identities, $\boldsymbol{q}_\ell$					$p(\boldsymbol{\lambda} \mid \text{data})$
	$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 4$	$\ell = 5$	mean & 90% CI
$\mathbf{1}_n$	0	0	0	0	0.004	0.000 (0.000, 0.000)
MKT	0	0	0	0	0.004	0.000 (0.000, 0.000)
MKT$\star$	1.000	0	0	0	0.008	0.314 (0.169, 0.461)
accruals	0	0	0	0	0.051	0.002 (-0.026, 0.029)
aturnover	0	0	0	0	0.060	0.020 (-0.004, 0.045)
betaarb	0	0	0	0	0.006	0.000 (-0.002, 0.001)
cfp	0	0	0	0	0.018	0.001 (-0.002, 0.004)
ciss	0	0	0	0	0.008	0.000 (-0.005, 0.004)
nissa	0	0	0	0	0.011	0.002 (-0.005, 0.010)
divp	0	0	0	0.001	0.011	-0.001 (-0.002, 0.001)
dur	0	0	0	0	0.021	0.001 (-0.002, 0.004)
ep	0	0	0	0	0.009	0.000 (-0.004, 0.005)
gmargins	0	0	0	0	0.018	-0.002 (-0.009, 0.005)
growth	0	0	0	0	0.015	0.002 (-0.004, 0.007)
igrowth	0	0	0	0	0.021	0.005 (-0.004, 0.015)
indmom	0	0	0	0	0.019	-0.001 (-0.004, 0.002)
indrrev	0	0	0	0	0.014	0.007 (0.002, 0.014)
inv	0	0	0	0	0.017	0.006 (-0.001, 0.013)
invcap	0	0	0	0	0.008	-0.001 (-0.004, 0.003)
lev	0	0	0	0	0.021	-0.001 (-0.006, 0.004)
lrrev	0	0	0	0	0.025	0.002 (-0.002, 0.006)
mom	0	0	0	0	0.018	0.000 (-0.003, 0.003)
mom12	0	0	0	0	0.015	0.001 (-0.001, 0.004)
momrev	0	0	0	0	0.033	0.006 (-0.002, 0.015)
noa	0	0	0	0	0.047	0.012 (-0.010, 0.035)
price	0	0	0	0	0.009	0.000 (-0.002, 0.001)
prof	0	0	0	0	0.031	0.005 (-0.006, 0.016)
roaa	0	0	0	0	0.008	0.000 (-0.005, 0.004)
roea	0	0	0	0	0.008	0.000 (-0.005, 0.004)
season	0	0	0	0.999	0.031	0.246 (0.127, 0.362)
sgrowth	0	0	0	0	0.013	0.000 (-0.006, 0.005)
shvol	0	0	0	0	0.006	-0.001 (-0.003, 0.002)
size	0	0	0	0	0.015	0.001 (-0.002, 0.004)
sp	0	0	0	0	0.022	0.001 (-0.002, 0.005)
strev	0	0	0	0	0.016	0.005 (0.000, 0.010)
valmom	0	0	0	0	0.018	0.000 (-0.002, 0.003)
valmomprof	0	0	0	0	0.020	0.002 (-0.001, 0.005)
valprof	0	1.000	0	0	0.024	0.348 (0.237, 0.458)
value	0	0	0	0	0.020	0.001 (-0.002, 0.003)
valuem	0	0	0	0	0.019	0.000 (-0.002, 0.002)
indmomrev	0	0	1.000	0	0.035	0.219 (0.098, 0.340)
age	0	0	0	0	0.007	0.001 (-0.004, 0.005)
debtiss	0	0	0	0	0.011	0.002 (-0.009, 0.014)
fscore	0	0	0	0	0.015	0.000 (-0.013, 0.013)
gltnoa	0	0	0	0	0.066	-0.001 (-0.055, 0.051)
indrrevlv	0	0	0	0	0.032	0.030 (0.012, 0.049)
nissm	0	0	0	0	0.011	0.001 (-0.006, 0.009)
ivol	0	0	0	0	0.006	0.000 (-0.004, 0.003)
repurch	0	0	0	0	0.010	0.000 (-0.011, 0.012)
roa	0	0	0	0	0.010	0.001 (-0.004, 0.005)
roe	0	0	0	0	0.010	0.001 (-0.004, 0.006)
rome	0	0	0	0	0.010	0.002 (-0.004, 0.009)
shortint	0	0	0	0	0.008	-0.002 (-0.009, 0.003)
sue	0	0	0	0	0.031	0.007 (-0.003, 0.016)
Ident. prob.	1.000	1.000	0.932	0.931	0.658	

This table presents the Bayesian estimates of the 51 long-short anomalies, with the control for two market factors. The data are at the monthly frequency, spanning from January 1974 to December 2019. In empirical applications, all variables are standardized to have unit variance per period. The data description is in Section 3.1. We use Algorithm 1, with $L = 10$ and $\sigma_0 = 0.5/\sqrt{12}$, to estimate the posterior factor probabilities within identities ($\{\boldsymbol{q}_\ell\}_{\ell=1}^L$). The last row reports the posterior identity probabilities ($\{\theta_\ell\}_{\ell=1}^L$).