Time-Varying Factor Premia: Evidence from US Long-Short Factor Portfolios, 1963–2025

Jack Allen, Arkyne Capital

July 2026

Abstract

Systematic equity managers commonly apply static factor weights when ranking stocks by multi-factor scores. This paper asks whether that assumption holds in the data. Using Fama–French long-short portfolio returns from July 1963 to December 2025 ($T = 750$ months), the paper treats the cross-sectional factor pricing model as a conditional model $E[r_{i,t+1} \mid \mathcal{F}_t] = \mathbf{f}_{i,t}^{\top}\lambda_t$ and tests the restriction $\lambda_t = \lambda$ implied by static weighting. A recursive expanding-window Newey–West mean serves as the restricted estimator; a 60-month rolling mean serves as the unrestricted alternative. Bai–Perron multiple structural break tests applied to monthly returns, with HAC inference, fail to reject the constant-mean null for any of the six factors considered, contradicting the strong rejections obtained when the same test is applied (improperly) to the rolling-mean series. A Diebold–Mariano test on 690 out-of-sample months at the baseline 60-month window delivers a consistent verdict: the rolling estimator does not improve predictive accuracy for the equal-weighted factor composite at any of the window lengths considered and is significantly worse than the static estimator for profitability (5%) and investment (10%), with no factor showing a significant improvement. The practical implication is that strategies which adjust factor weights in response to rolling-premium estimates are responding to statistical noise rather than genuine regime shifts—and incur measurable cost in doing so.

1 Introduction

The multi-factor model is the dominant framework in systematic equity investing. Managers score stocks on characteristics such as value, momentum, and profitability, then tilt portfolios toward stocks with favourable composite scores. In its standard implementation, the weight assigned to each factor is estimated over a long historical sample and held fixed thereafter. This static approach has a clean theoretical foundation in the Fama–MacBeth (Fama and MacBeth 1973) two-pass cross-sectional regression and a large empirical literature establishing that factor premia are positive and economically meaningful on average.

What the static approach implicitly assumes is that the return per unit of factor exposure is approximately constant over time—holding one unit of value exposure today earns roughly the same expected premium as it did a decade ago. This assumption is directly challenged by the conditional asset pricing literature. Ferson and Harvey (1991) show that expected returns vary with observable state variables; Cochrane (1996) argues that risk prices are functions of the business cycle; Lettau and Ludvigson (2001) demonstrate that the equity premium co-moves with the consumption-to-wealth ratio. At the individual factor level the evidence is more direct: the value premium effectively vanished over the 2010s (Fama and French 2020), momentum crashed in 2009 and has been highly volatile since (Daniel and Moskowitz 2016), and the size premium has been near zero since the early 1980s once risk adjustments are applied (Hou, Xue, and Zhang 2020). A practitioner observing these patterns would have strong intuitive grounds for adjusting factor weights over time. The question this paper poses is whether that intuition is statistically justified.

This paper formalises the question by writing the cross-sectional factor pricing model in conditional form, $E[r_{i,t+1} \mid \mathcal{F}_t] = \mathbf{f}_{i,t}^{\top}\lambda_t$, and treating the static-weighting assumption as the testable restriction $\lambda_t = \lambda$ for all t. Rather than modelling λ_t as a parametric function of state variables, as in the conditional pricing literature (Ferson and Harvey 1991; Cochrane 1996; Lettau and Ludvigson 2001), the approach taken here treats λ_t as an unspecified time-varying object and asks whether its mean is constant over time, using rolling-window estimates and structural break tests on long-short portfolio returns. The empirical design compares two nested estimators: a full-sample Newey–West mean (the restricted estimator that imposes constant premia) and a 60-month rolling mean (the unrestricted alternative that allows premia to drift).

The two estimators are evaluated in two ways: structural break tests on monthly returns to detect regime shifts, and Diebold–Mariano predictive accuracy tests on the equal-weighted factor composite to assess out-of-sample consequences.

The paper is organised as follows. Section 2 presents the theoretical framework. Section 3 describes the data. Section 4 sets out the estimators and formal tests. Section 5 reports the empirical results. Section 6 discusses the practical implications of the findings. Section 7 concludes.

2 Framework

2.1 Factor Pricing and Cross-Sectional Regression

The starting point is the standard linear factor pricing model in conditional form. Expected excess returns are linear in factor exposures, with potentially time-varying premia:

$$E[r_{i,t+1} \mid \mathcal{F}_t] = \mathbf{f}_{i,t}^{\top} \lambda_t$$

where $r_{i,t+1}$ is the excess return of stock i from t to $t+1$, $\mathbf{f}_{i,t} \in \mathbb{R}^K$ is the vector of firm-level factor exposures at time t, and $\lambda_t \in \mathbb{R}^K$ is the vector of factor premia—the return per unit of factor exposure that investors require as compensation for systematic risk in the conditioning environment $\mathcal{F}_t$.

The associated cross-sectional pricing equation is:

$$r_{i,t+1} = \alpha_t + \mathbf{f}_{i,t}^{\top} \lambda_t + \varepsilon_{i,t+1}, \qquad E[\varepsilon_{i,t+1} \mid \mathbf{f}_{i,t}] = 0$$

The static factor weighting that dominates practice is equivalent to the restriction $\lambda_t = \lambda$ for all t. Whether this restriction holds is an empirical question; the remainder of the paper provides a formal answer.

2.2 Estimating Premia from Factor Portfolios

The cleanest implementation uses pre-formed long-short factor portfolios rather than stock-level cross-sections. When the long-short portfolio is correctly constructed, its

expected return equals the factor premium directly: $E[r^{\text{LS}}_{k,t} \mid \mathcal{F}_{t-1}] = \lambda_{k,t}$.[1] Under the static restriction, the full-sample estimate is the time-series mean:

$$\hat{\lambda} = \frac{1}{T}\sum_{t=1}^{T} r^{\text{LS}}_{t}$$

with a Newey–West standard error to account for serial correlation in monthly returns. This approach bypasses the two-pass Fama–MacBeth procedure—and the associated errors-in-variables problem identified by Shanken (1992)—because no first-pass beta estimation is required. The factor betas are embedded in the portfolio construction, not estimated separately.

2.3 The Time-Varying Alternative

If premia genuinely vary, the full-sample average conflates structurally distinct regimes. The natural relaxation is to estimate each premium using only the W most recent months:

$$\hat{\lambda}^{\text{Roll}}_{k,t}(W) = \frac{1}{W}\sum_{s=t-W+1}^{t} r^{\text{LS}}_{k,s}$$

for each factor k. The window length W governs a bias-variance tradeoff: short windows adapt quickly to regime changes but are noisy; long windows are precise but slow to respond. The baseline is $W = 60$ months (five years), with $W \in \{36, 84\}$ as robustness checks.[2] The static estimator is nested in the rolling estimator as the limiting case $W = T$.

[1] The long-short portfolio for factor k is constructed to have unit exposure to factor k and zero exposure to all other factors (Fama and French 1993). Writing the portfolio return as $r^{\text{LS}}_{k,t+1} = \sum_i w^{\text{LS}}_{i,t} r_{i,t+1}$ with weights satisfying $\sum_i w^{\text{LS}}_{i,t} f_{i,k',t} = \mathbf{1}\{k' = k\}$ for all k', and substituting the conditional pricing equation, gives $E[r^{\text{LS}}_{k,t+1} \mid \mathcal{F}_t] = \lambda_{k,t}$, since contributions from all factors $k' \neq k$ net out across the long and short legs.

[2] Sixty months is a standard window length in the empirical asset pricing literature—long enough that the conditional mean is estimated with reasonable precision (a Newey–West standard error of order $\sigma/\sqrt{60}$ on monthly returns) but short enough to track regime variation over the business cycle (Lewellen, Nagel, and Shanken 2010). Section 5.2 reports the rolling-premium plot for all three window lengths; the choice of W affects smoothness but not the qualitative pattern.

Table 1: Summary statistics for the six Fama–French long-short factor portfolios, July 1963–December 2025. Returns are monthly. The t-statistic uses Newey–West standard errors (6 lags). Ann. Sharpe is annualised ($\times\sqrt{12}$). Negative skewness indicates left-tail risk.

Factor	Mean (%)	SD (%)	t-stat	Ann. Sharpe	Skewness	Ex. Kurtosis
Investment (CMA)	0.24	2.06	3.20	0.40	0.27	1.36
Momentum (UMD)	0.59	4.17	3.90	0.49	-1.30	9.74
Profitability (RMW)	0.26	2.22	3.23	0.41	-0.32	10.96
Short-term Reversal (STR)	0.41	3.15	3.55	0.45	0.44	5.50
Size (SMB)	0.18	3.03	1.62	0.20	0.36	3.03
Value (HML)	0.29	2.97	2.65	0.34	0.10	2.17

3 Data

The analysis uses monthly returns on six long-short factor portfolios from Kenneth French's publicly available data library: SMB (size), HML (value), RMW (profitability), and CMA (investment) from the Fama–French five-factor model (Fama and French 2015); UMD (momentum) (Jegadeesh and Titman 1993); and STR (short-term reversal) (Jegadeesh 1990). The sample runs from July 1963 to December 2025, giving $T = 750$ monthly observations per factor. Returns are net of the risk-free rate and incorporate delisting returns according to the standard convention used by the French data library.

Table 1 reports summary statistics for the six factor portfolios. All six carry positive unconditional means; momentum, profitability, short-term reversal, and investment are significant at the 1% level under Newey–West standard errors. The factor portfolios exhibit pronounced negative skew and excess kurtosis in momentum and profitability, consistent with the crash dynamics documented by Daniel and Moskowitz (2016). Figure 1 plots cumulative log returns over the six-decade sample. Each panel shows extended periods where the cumulative path flattens or reverses—the visual case for time-variation that the formal tests in Section 5 evaluate.

4 Methodology

The sample contains $T = 750$ monthly returns (July 1963 to December 2025). After a W-month burn-in, the rolling estimator produces $T - W$ observations per factor. The out-of-sample evaluation window runs from month $W+1$ to T, giving $n = 714, 690, 666$ observations for $W = 36, 60, 84$ respectively.

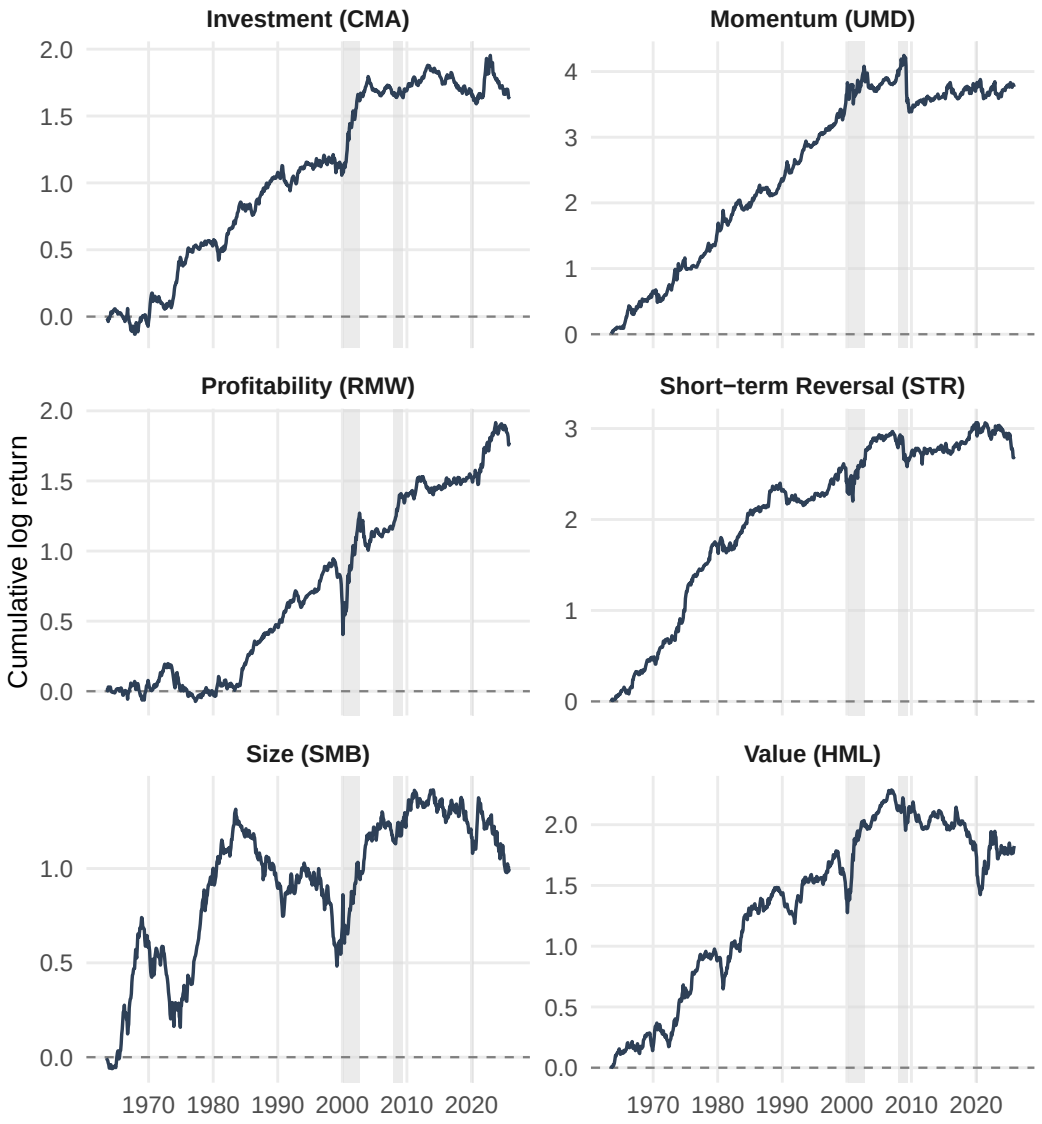

Figure 1: Cumulative log returns of the six Fama–French long-short factor portfolios, July 1963–December 2025. Shaded regions mark the dot-com crash (2000–2002), the global financial crisis (2008–2009), and the COVID shock (2020).

Static Newey–West mean. The full-sample estimator $\hat{\lambda} = T^{-1}\sum_t r_t^{\mathrm{LS}}$ is reported on the full 1963–2025 sample (with Newey–West standard errors at 6 lags) to characterise each factor's unconditional premium. For the out-of-sample comparison, the same estimator is implemented recursively: at each out-of-sample month t, $\hat{\lambda}_{k,t}^{\mathrm{Stat}} = t^{-1}\sum_{s=1}^{t} r_{k,s}^{\mathrm{LS}}$ is computed using only data available up to time t, avoiding look-ahead bias. This is the implicit procedure of any manager who sets factor weights proportional to historical premia and updates them only as the sample grows.

Rolling mean. For each month t after a $W = 60$ month burn-in, the rolling estimator computes $\hat{\lambda}_{k,t}^{\mathrm{Roll}}(W)$ using only the preceding W months. Window-length sensitivity is reported for $W \in \{36, 60, 84\}$.

Bai–Perron structural break test. The test of $\lambda_t = \lambda$ is formalised as a multiple structural breaks problem in the mean of each long-short return series, following Bai and Perron (1998) and Bai and Perron (2003). The unrestricted model partitions the sample into $m+1$ regimes separated by m unknown break dates $\tau_1 < \cdots < \tau_m$:

$$r_{k,t}^{\mathrm{LS}} = \mu_k^{(j)} + \varepsilon_{k,t}, \qquad t \in (\tau_{j-1}, \tau_j], \quad j = 1, \ldots, m+1. \tag{1}$$

Break dates and segment means are estimated jointly by minimising the sum of squared residuals over admissible partitions, subject to a trimming requirement that each segment contain at least $\lfloor \epsilon T \rfloor$ observations ($\epsilon = 0.15$). HAC inference uses Newey–West weights at 6 lags. The final number of breaks $\hat{m}$ is selected by the Bayesian information criterion:

$$\mathrm{BIC}(m) = \log \hat{\sigma}^2(m) + \frac{p_m \log T}{T}, \tag{2}$$

where $p_m = 2m + 1$ counts free parameters. Applying the test directly to monthly returns is essential: applying it to the rolling-mean series introduces construction-induced autocorrelation (adjacent observations of a 60-month rolling mean share 59 of 60 underlying returns), under which standard sup-F critical values over-reject and the test loses its interpretation as evidence on λ_t.

Diebold–Mariano predictive accuracy test. The out-of-sample predictive accuracy of the static and rolling estimators is compared using the test of Diebold and Mariano (1995), with the small-sample correction of Harvey, Leybourne, and Newbold

Table 2: Full-sample factor premia: time-series means of monthly Fama–French long-short portfolio returns, July 1963–December 2025. $\hat{\lambda}$ is the sample mean in percent per month. The Rolling avg column reports the time-series average of the 60-month rolling estimate over the out-of-sample evaluation period (1968–2025). Standard errors are Newey–West with 6 lags. Significance: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Factor	Static $\hat{\lambda}$ (%)	SE (NW)	t-stat	Rolling avg (%)
Investment (CMA)	0.241	0.086	2.79***	0.276
Momentum (UMD)	0.594	0.155	3.84***	0.597
Profitability (RMW)	0.262	0.090	2.92***	0.281
Short-term Reversal (STR)	0.409	0.108	3.80***	0.455
Size (SMB)	0.179	0.118	1.52	0.192
Value (HML)	0.288	0.129	2.23**	0.282

(1997). At each out-of-sample month t, each estimator generates a one-step-ahead prediction of the equal-weighted factor composite return $y_{t+1} = K^{-1} \sum_k r^{\mathrm{LS}}_{k,t+1}$. Under squared-error loss the per-period loss differential is

$$d_t = e^2_{\mathrm{Stat},t} - e^2_{\mathrm{Roll},t}, \tag{3}$$

with the DM statistic

$$\mathrm{DM} = \frac{\bar{d}}{\sqrt{\hat{\sigma}^2(\bar{d})/n}}, \tag{4}$$

where $\hat{\sigma}^2(\bar{d})$ is a Newey–West HAC estimator with bandwidth set by the automatic data-dependent rule of Newey and West (1994). A positive DM statistic indicates lower mean squared error for the rolling estimator; a negative DM indicates the reverse.

5 Empirical Results

5.1 Stage 1: Static Premia

Before testing whether premia vary, the static estimator establishes that they exist at all. Table 2 reports the full-sample Newey–West mean of each factor's long-short return over the 1963–2025 sample.

All six factors carry positive unconditional premia over the full sample, and most

are significant at the 1% level. The close correspondence between the Static $\hat{\lambda}$ and Rolling avg columns confirms that both estimators target the same underlying quantity. The empirical question is whether the rolling series is stable around that average, or whether it makes large excursions that the long-run mean conceals.

5.2 Stage 2: Time-Varying Premia

Figure 2 plots the rolling premium series for each factor across the three window lengths. The series are visibly non-constant: HML swings from strongly positive in the 1970s and early 2000s to near zero or negative across the 2010s before recovering post-2022, while UMD spends extended periods in negative territory around the 2009 crash (Daniel and Moskowitz 2016). The three window lengths track closely and differ mainly in smoothness, indicating the pattern is not an artefact of bandwidth choice. The visible variation looks compelling, but as the formal tests below show, sampling variability in rolling-window estimates of noisy monthly returns mechanically generates trajectories of this kind: visual time-variation is not the same as statistically detectable time-variation.

Table 3 formalises the visual evidence with Bai–Perron structural break tests applied directly to the monthly long-short return series, with HAC inference and BIC selecting the number of breaks from a maximum of five. The BIC-optimal number of breaks is zero for every factor. The Gap column shows by how much: even the best single-break partition delivers fit improvements smaller than the per-break BIC penalty of $2 \log T \approx 13.24$, with gaps ranging from 3.86 (UMD) to 8.22 (RMW). The two factors with the smallest gaps—momentum and short-term reversal—are precisely those where the Andrews (1993) sup-Wald statistic against a single-break alternative does reject (supF = 8.06, 10% for UMD; supF = 11.10, 5% for STR). For the remaining four factors, the visual variation in Figure 2 cannot be distinguished from sampling noise once the construction-induced autocorrelation of the rolling-mean series is removed. This pattern is substantive: rejection for at most two of six factors—and only under a single-break alternative—implies that earlier evidence of regime shifts in factor premia, often presented using rolling estimators, may overstate the formal statistical case for time-variation.[3]

[3]As a size check, the empirical false-positive rate is simulated under the constant-mean null. For each factor, the monthly return series is demeaned, residuals are resampled via the stationary bootstrap of Politis and Romano (1994) with mean block length 12, and the full-sample mean is added back, giving $B = 1{,}000$ replications with constant true mean by construction. Applying the

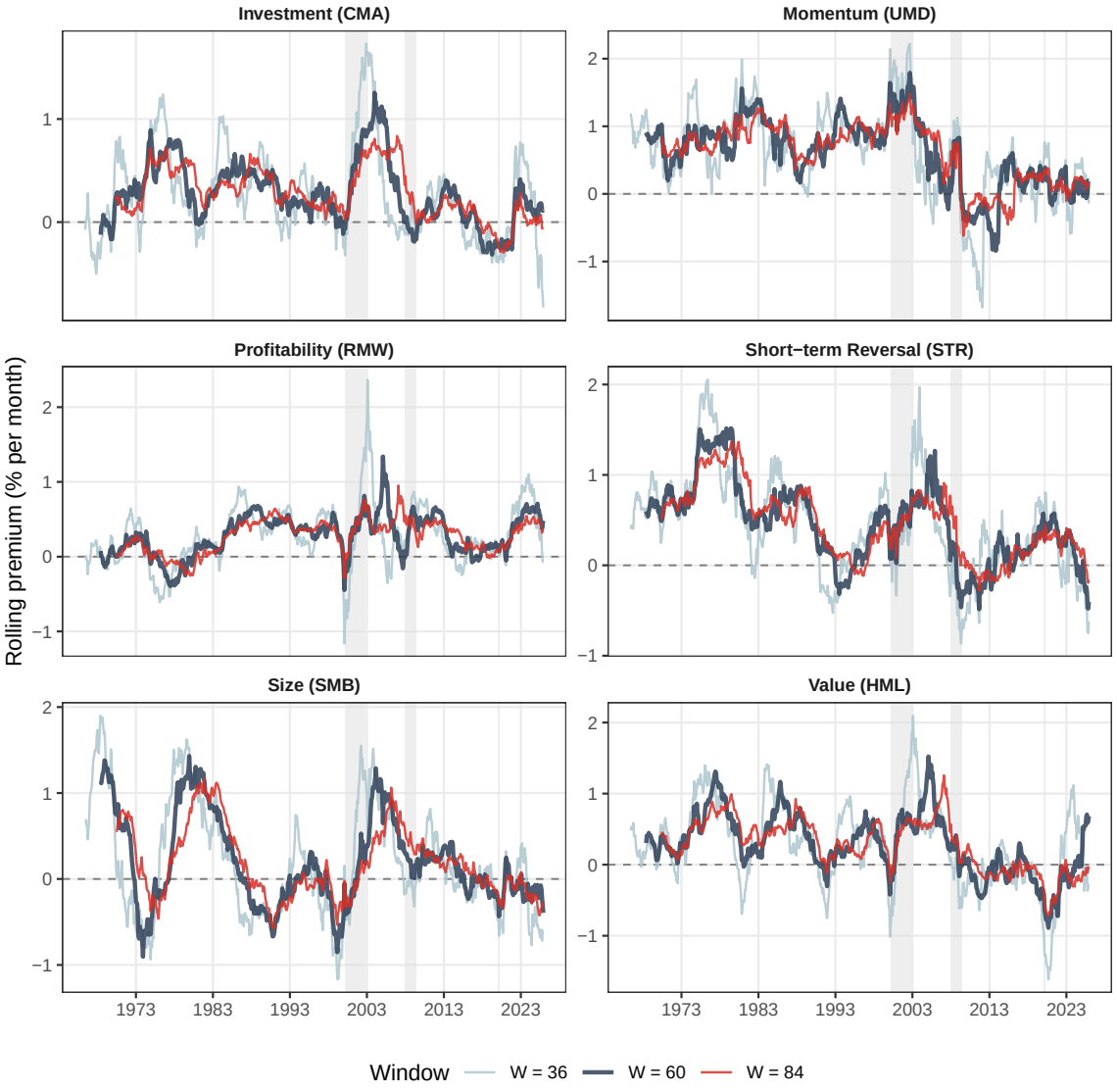

Figure 2: Time-varying factor premia: rolling means of Fama–French long-short portfolio returns (%, per month), July 1963–December 2025. Each panel shows one factor across three window lengths: $W = 36$ (light), $W = 60$ (medium, baseline), and $W = 84$ (dark). The dashed line marks zero. Grey shading marks the dot-com crash (2000–2002), global financial crisis (2008–2009), and COVID shock (2020).

Table 3: Bai–Perron structural break tests applied to monthly Fama–French long-short returns, July 1963–December 2025 ($T = 750$ months per factor). The BIC-optimal number of breaks $\hat{m}$ is selected from a maximum of five; the result is unchanged when the cap is raised to six. ΔFit at $m = 1$ is the change in $T \log(\hat{\sigma}^2)$ between the constant-mean model and the BIC-optimal single-break partition. BIC penalty is the per-break BIC penalty $2 \log T$; Gap = BIC penalty $-$ $|\Delta\text{Fit}|$ (positive values keep $\hat{m} = 0$). supF (HAC) is the Andrews (1993) sup-Wald statistic against a single-break alternative; critical values are 7.17 (10%), 8.85 (5%), 12.35 (1%). Significance: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Factor	$\hat{m}$	ΔFit at $m = 1$	BIC penalty	Gap	supF (HAC)
Investment (CMA)	0	-7.50	13.24	5.74	6.36
Momentum (UMD)	0	-9.25	13.24	3.99	7.92*
Profitability (RMW)	0	-5.01	13.24	8.23	4.86
Short-term Reversal (STR)	0	-8.70	13.24	4.54	11.25**
Size (SMB)	0	-6.80	13.24	6.44	5.18
Value (HML)	0	-7.11	13.24	6.13	4.80

5.3 Stage 3: Predictive Accuracy

The break test in Section 5.2 does not reject parameter stability at the factor level. Whether the unrestricted estimator nonetheless improves out-of-sample prediction is a logically separate question with direct practical relevance: a manager only cares about time-variation if exploiting it improves their forecasts. This is the cross-sectional analogue of the equity-premium-prediction question studied by Welch and Goyal (2008), who show that the historical mean consistently outperforms a wide range of conditional predictors out of sample, even when those predictors have in-sample explanatory power. The same question is posed here for factor premia.

Table 4 reports Diebold–Mariano tests at the baseline $W = 60$, broken out factor-by-factor with the equal-weighted composite as the final row. Every per-factor DM statistic is negative, indicating the rolling estimator has weakly higher mean squared error than the recursive static estimator at the factor level; the rejection is significant at the 5% level for RMW (DM $= -2.26$) and at the 10% level for CMA (DM $= -1.89$). The composite DM is positive but statistically indistinguishable from zero. The table delivers a stronger verdict than mere failure to reject: the rolling estimator does

same BP–BIC specification yields empirical false-positive rates of 3.3% (SMB), 4.6% (HML), 5.4% (RMW), 5.1% (CMA), 0.5% (UMD), and 0.5% (STR). Each is in the neighbourhood of the nominal 5%, consistent with a well-sized procedure under the null. The failure to reject on actual returns is therefore not an artefact of an over-rejecting test.

Table 4: Diebold–Mariano predictive accuracy tests at the baseline rolling window $W = 60$ months, comparing the recursive expanding-window static estimator against the rolling-mean estimator. All tests use $T_{\text{OOS}} = 690$ months (1968–2025). Each factor row tests predictive accuracy for the realised factor return $r_{k,t+1}$; the composite row tests for the equal-weighted composite $y_{t+1} = K^{-1} \sum_k r_{k,t+1}$. MSE columns are in units of 10^{-4}. A positive DM statistic indicates the static estimator has higher MSE (rolling preferred); a negative DM indicates the reverse. The Harvey–Leybourne–Newbold small-sample correction is applied. Significance: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Target	MSE (static) $\times 10^4$	MSE (rolling) $\times 10^4$	DM statistic	p-value
Size (SMB)	9.263	9.363	-0.844	0.398
Value (HML)	9.340	9.451	-0.831	0.406
Profitability (RMW)	5.186	5.287	-2.260**	0.024
Investment (CMA)	4.276	4.373	-1.896*	0.058
Momentum (UMD)	18.278	18.480	-1.564	0.118
Short-term Reversal (STR)	10.512	10.521	-0.103	0.918
Equal-weighted composite	1.340	1.325	0.593	0.553

not improve predictive accuracy on any factor, and is significantly worse on two. Window-length sensitivity is contained—composite DM statistics at $W \in \{36, 84\}$ are -0.14 and 0.45 (untabulated), neither significant.

6 Practical Implications

The findings in Section 5 carry a specific implication for investment practice that is worth drawing out directly. A multi-factor equity strategy involves, at minimum, two decisions: which factors to hold, and how much weight to assign to each. The existing literature has concentrated almost entirely on the first question—establishing which factors are robustly priced—while treating the second as a static calibration problem solved once over a long backtest. The present results suggest that this ordering is empirically justified.

The rolling premium series in Figure 2 present what appears, visually, to be a compelling case for dynamic reweighting. Value premia were large in the 1970s and around the early 2000s recovery, then negative for much of the 2010s. Momentum crashed sharply in 2009. A discretionary manager observing these patterns in real time would have strong intuitive grounds for adjusting allocations: increasing weight on factors whose recent premia appear elevated and reducing weight on those where

the rolling estimate has turned negative. The formal tests in Sections 5.2 and 5.3 provide a precise statistical answer to whether that intuition has empirical support. It does not.

The Bai–Perron tests show that the variation visible in Figure 2 is statistically indistinguishable from the sampling variability expected under a constant-mean null. A 60-month rolling window applied to a series with monthly return volatility of 3–5% produces trajectories that look highly non-stationary even when the underlying mean is fixed. The construction-induced persistence of the rolling-mean series amplifies this visual impression: adjacent observations share 59 of 60 underlying data points, so any short-run deviation from the long-run mean propagates smoothly across the window, creating the appearance of extended regime shifts where none exist in the underlying data. The BIC selection of zero breaks for all six factors, with substantial gaps between the fit improvement at the best single-break partition and the penalty required to justify that break, indicates that the data contain no break large enough to overcome the cost of the additional parameters.

The Diebold–Mariano results in Table 4 are the more direct investment-relevant finding. Even setting aside the structural break evidence, one might argue that a rolling estimator can add value through gradual adaptation to slow-moving changes in premia that are too diffuse to register as discrete breaks. The DM test evaluates this directly: over 690 out-of-sample months, does the rolling estimator produce lower mean squared error for realised factor returns than the recursive expanding-window mean? For every factor, the answer is no, and for two factors the rolling estimator is significantly worse. This places factor premium prediction in the same empirical category as equity premium prediction in Welch and Goyal (2008): the historical mean—the simplest estimator consistent with constant premia—is not only as good as the more flexible alternative, it is better. Adding flexibility that the data do not support inflates estimation variance without reducing bias, producing a net cost.

The implication for systematic versus discretionary approaches to factor allocation is specific. A systematic strategy that computes factor scores, weights them by long-run historical premia, and rebalances mechanically according to a fixed rule holds factor weights that are, on this evidence, correctly specified for the population distribution of returns. The regularity with which the strategy applies that rule—regardless of what the rolling premium series appears to show at any point—is not merely a process

virtue; it is empirically appropriate given the failure of the unrestricted estimator to improve out-of-sample prediction. A discretionary approach that varies factor weights in response to recent rolling estimates, by contrast, is responding to noise and incurring the additional estimation error captured by the negative DM statistics for RMW and CMA. This does not imply that factor premia are constant or that time-variation is theoretically impossible. The power caveat in Section 5.2 is genuine: break magnitudes small relative to monthly return volatility may exist without being detectable on $T = 750$ observations. What the results do establish is that, over the six-decade sample examined, no adjustment to static factor weights based on rolling-window premium estimates would have improved predictive accuracy—and two adjustments would have measurably reduced it.

7 Conclusion

This paper tests the restriction $\lambda_t = \lambda$ implied by static factor weighting in systematic equity management, using Fama–French long-short portfolio returns from July 1963 to December 2025. Two lines of formal evidence deliver a consistent verdict.

First, Bai–Perron multiple structural break tests applied to monthly long-short returns with HAC inference fail to reject the constant-mean null for any of the six factors; BIC selects zero breaks for every factor under both five- and six-break caps. A stationary-bootstrap size check confirms that the BP–BIC pipeline is well-sized to slightly conservative under the constant-mean null, so the absence of rejection is not an artefact of an over-rejecting test.

Second, the Diebold–Mariano test on the equal-weighted factor composite fails to detect a difference in out-of-sample predictive accuracy between the recursive expanding-window static estimator and the 60-month rolling mean at any of the three window lengths considered. At the factor level the rolling estimator is significantly worse for RMW (5%) and CMA (10%), with no factor showing a significant improvement. The visible time-variation in Figure 2, which would lead a practitioner to update factor weights if observed in real time, does not survive formal econometric testing once the construction-induced autocorrelation of the rolling series is removed.

This null result is itself substantive. A literature that documents apparent regime shifts in factor premia using rolling-window estimators—often without inference on

the underlying monthly returns—may overstate the formal statistical case for time-variation. Such evidence as exists, on the present sample, is concentrated in short-term reversal and (more weakly) momentum, the two factors with the most pronounced trading-frequency behaviour. Different estimators (state-space models, predictive-systems frameworks), different conditioning information, or longer samples could yield different inferences. The paper characterises what these tests can detect on this data; it does not foreclose the possibility of time-variation outside their reach.

References

Andrews, Donald W. K. 1993. "Tests for Parameter Instability and Structural Change with Unknown Change Point." *Econometrica* 61 (4): 821–56.

Bai, Jushan, and Pierre Perron. 1998. "Estimating and Testing Linear Models with Multiple Structural Changes." *Econometrica* 66 (1): 47–78.

———. 2003. "Computation and Analysis of Multiple Structural Change Models." *Journal of Applied Econometrics* 18 (1): 1–22.

Cochrane, John H. 1996. "A Cross-Sectional Test of an Investment-Based Asset Pricing Model." *Journal of Political Economy* 104 (3): 572–621.

Daniel, Kent, and Tobias J. Moskowitz. 2016. "Momentum Crashes." *Journal of Financial Economics* 122 (2): 221–47.

Diebold, Francis X., and Roberto S. Mariano. 1995. "Comparing Predictive Accuracy." *Journal of Business and Economic Statistics* 13 (3): 253–63.

Fama, Eugene F., and Kenneth R. French. 1993. "Common Risk Factors in the Returns on Stocks and Bonds." *Journal of Financial Economics* 33 (1): 3–56.

———. 2015. "A Five-Factor Asset Pricing Model." *Journal of Financial Economics* 116 (1): 1–22.

———. 2020. "Comparing Cross-Section and Time-Series Factor Models." *Review of Financial Studies* 33 (5): 1891–1926.

Fama, Eugene F., and James D. MacBeth. 1973. "Risk, Return, and Equilibrium: Empirical Tests." *Journal of Political Economy* 81 (3): 607–36.

Ferson, Wayne E., and Campbell R. Harvey. 1991. "Variation in Expected Security Returns, Risk Premiums, and the Test of Asset Pricing Models." *Journal of Finance* 46 (2): 663–91.

Harvey, David, Stephen Leybourne, and Paul Newbold. 1997. "Testing the Equality

of Prediction Mean Squared Errors." *International Journal of Forecasting* 13 (2): 281–91.

Hou, Kewei, Chen Xue, and Lu Zhang. 2020. "Replicating Anomalies." *Review of Financial Studies* 33 (5): 2019–2133.

Jegadeesh, Narasimhan. 1990. "Evidence of Predictable Behavior of Security Returns." *Journal of Finance* 45 (3): 881–98.

Jegadeesh, Narasimhan, and Sheridan Titman. 1993. "Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency." *Journal of Finance* 48 (1): 65–91.

Lettau, Martin, and Sydney Ludvigson. 2001. "Resurrecting the (c)CAPM: A Cross-Sectional Test When Risk Premia Are Time-Varying." *Journal of Political Economy* 109 (6): 1238–87.

Lewellen, Jonathan, Stefan Nagel, and Jay Shanken. 2010. "A Skeptical Appraisal of Asset-Pricing Tests." *Journal of Financial Economics* 96 (2): 175–94.

Newey, Whitney K., and Kenneth D. West. 1994. "Automatic Lag Selection in Covariance Matrix Estimation." *Review of Economic Studies* 61 (4): 631–53.

Politis, Dimitris N., and Joseph P. Romano. 1994. "The Stationary Bootstrap." *Journal of the American Statistical Association* 89 (428): 1303–13.

Shanken, Jay. 1992. "On the Estimation of Beta-Pricing Models." *Review of Financial Studies* 5 (1): 1–33.

Welch, Ivo, and Amit Goyal. 2008. "A Comprehensive Look at the Empirical Performance of Equity Premium Prediction." *Review of Financial Studies* 21 (4): 1455–1508.

Disclaimer

This paper is published by Arkyne Capital for informational and research purposes only. It does not constitute financial advice, an offer to invest, or a recommendation to buy or sell any financial product. The empirical analysis uses publicly available data from Kenneth French's data library and relates to US equity factor portfolios; results may not apply in other markets, time periods, or investment contexts. Past performance and historical return patterns are not indicative of future results. Arkyne Capital makes no representations as to the accuracy, completeness, or suitability of the information herein for any particular purpose. This document is intended for sophisticated readers with relevant financial knowledge and experience. It should not be relied upon as the sole basis for any investment decision.