

Market Regimes and Model Uncertainty in Portfolio Construction: Select or Diversify?

THIS VERSION: SEPTEMBER 27, 2026

Mario Bajo Traver

Mario Bajo Traver is Head of Investment Strategies Unit, Department of Market Analysis and Intelligence, Banco de España. mario.bajo@bde.es

KEY FINDINGS

- The performance ranking of portfolio construction methods reorders across market states: from calm to stress, maximum-Sharpe, the most estimation-dependent method, falls from the strongest to the weakest performer, while equal weighting rises from the weakest to the strongest, so which method to trust is state-dependent.
- The market state predicts which method will perform better beyond the methods' own recent performance. The signal is concentrated in the most estimation-dependent method, survives placebo tests and a multiplicity correction, and persists, though weaker, when expected returns are shrunk.
- Exploiting the signal produces economic value net of costs: a state-conditional selector significantly outperforms performance-chasing, while its edge over diversification is positive but not significant. The advantage over performance-chasing tends to be larger when methods disagree.

ABSTRACT

Building a portfolio requires choosing a construction method (minimum-variance, maximum-Sharpe, risk parity, or a hierarchical method) whose relative performance shifts over time, exposing the investor to model uncertainty beyond the parameter uncertainty these methods are designed to address. We ask whether the observable market state carries information about which method will perform better next period, beyond the methods' own recent performance. Framing the test as one of conditional predictive ability, we find that it does: a state defined by realized volatility predicts relative method performance, with the effect concentrated in the most estimation-dependent method, consistent with an estimation-stress channel. The signal survives permutation, wrong-mapping, and strong-benchmark placebos. Exploiting it yields economic value net of costs: a state-conditional selector significantly outperforms performance-chasing, while its advantage over diversification is positive but not significant. The case for selection over performance-chasing tends to be stronger when methods disagree more. Model uncertainty at the construction stage is, in part, predictable from the market state.

JEL Codes: G11, G17, C58, C53

Keywords: Portfolio construction, Model uncertainty, Market regimes, Model selection, Conditional predictive ability, Estimation error, Naive diversification

DISCLOSURE

The author reports no potential conflict of interest. The opinions and analyses presented are the author's own and do not necessarily reflect those of Banco de España or the Eurosystem. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

An investor who builds a portfolio must first decide how to build it. Minimum-variance, maximum-Sharpe, risk parity, and hierarchical methods such as Hierarchical Risk Parity impose different assumptions on expected returns, covariances, and the dependence structure among assets, and they produce materially different allocations from the same data. The natural question, *which method is best*, assumes that the answer is stable. It is not: a method that dominates in calm markets can be among the worst when volatility rises, which is precisely when the estimated inputs it relies on are least reliable. Choosing among portfolio construction methods is therefore a decision under model uncertainty, and the cost of getting it wrong is concentrated in the environments where portfolios matter most.

Two responses to that uncertainty are available. An investor can select a single method and accept the risk that it is wrong for the prevailing environment, or diversify across methods by averaging their portfolios. The evidence on both is almost entirely unconditional: simple rules, above all 1/N, are difficult to beat out of sample (DeMiguel, Garlappi, and Uppal 2009), and combining rules can improve on their components (Tu and Zhou 2011), but neither result asks when one response should be preferred. What no study establishes is whether the market state carries information about which construction method will perform better next period, beyond what the methods' own recent performance already reveals, or when selecting among methods is worth more than diversifying across them. Without that evidence, an investor either commits to one method blindly or diversifies unconditionally, and in both cases leaves state information unused.

Contribution of this work. This paper establishes that the market state does carry such information. We treat the choice of portfolio construction method as a state-dependent decision and test it as a set of falsifiable hypotheses on a multi-asset universe of twelve ETFs from 2007 to 2026, using a walk-forward out-of-sample design. The pivotal test asks whether a realized-volatility state predicts the next-period performance differential between two methods after conditioning on their recent relative performance, a conditional-predictive-ability test in the sense of Giacomini and White (2006).

We document three results. *First, the performance ranking of methods reorders sharply across states*: maximum-Sharpe, the most estimation-dependent method, earns an annualized Sharpe ratio of 1.12 in calm markets and -0.35 in stress, while equal weighting moves in the opposite direction. *Second, the state predicts which method will do better beyond recent performance*: the test rejects for every pair involving maximum-Sharpe, and the result survives three placebo tests¹, a multiplicity correction, expected-return shrinkage, and leave-one-year-out tests. *Third, exploiting the information pays*: a state-conditional selector has the highest point estimate of the out-of-sample Sharpe ratio and, net of transaction costs, significantly outperforms performance-chasing; its edge over diversification is positive but not significant. The case for selection over performance-chasing tends to strengthen with economically meaningful cross-model disagreement. The implication is that model uncertainty at the construction stage is, in part, predictable from observable conditions.

The finding matters for how portfolios are built in practice. It rests on a simple, observable signal rather than a new construction method or a complex regime detector:

¹ Permutation, wrong-mapping, and strong-benchmark placebos test, respectively, whether the result depends on the temporal state–outcome alignment, the economically motivated state-to-method mapping, or the choice of benchmark against which incremental predictive content is assessed.

the prevailing volatility state carries information about when the most estimation-dependent method becomes less reliable, beyond what its own recent performance reveals. The information is primarily defensive, and cross-model disagreement provides a rough guide to when the choice of method matters enough to act on: when methods agree, diversification remains an efficient default.

Estimation error and model uncertainty. Optimized portfolios are fragile because their inputs are estimated. Mean-variance weights (Markowitz 1952) amplify estimation noise, and the damage is concentrated in expected returns rather than in the covariance matrix: errors in means are an order of magnitude more costly than errors in variances (Chopra and Ziemba 1993; Kan and Zhou 2007). The remedies follow the same asymmetry. Bayes–Stein shrinkage pulls sample means toward the minimum-variance portfolio's mean (Jorion 1986), while covariance shrinkage (Ledoit and Wolf 2004) and, revealingly, “wrong” portfolio constraints (Jagannathan and Ma 2003) improve out-of-sample risk by suppressing the extreme positions that estimation error would otherwise produce. All of these address parameter uncertainty, the estimated inputs of a given construction method. Model uncertainty, which method to use at all, is treated with care upstream, over return models, through Bayesian priors and multi-prior robustness (Pástor 2000; Garlappi, Uppal, and Wang 2007; Anderson and Cheng 2022); at the construction stage it is handled implicitly, by picking one method and living with it or by robustifying its inputs. Neither literature asks whether the *reliability* of a construction method itself varies with observable conditions.

The construction menu. The methods compared here differ in how heavily they lean on estimated inputs, and that ordering is the backbone of the mechanism. At the demanding end, maximum-Sharpe requires both expected returns and the inverse covariance matrix; minimum-variance and maximum-diversification (Choueifaty and Coignard 2008; Clarke, de Silva, and Thorley 2013) require the covariance or the correlation matrix but not expected returns; the hierarchical methods (López de Prado 2016; Raffinot 2018a, 2018b) use the covariance, through a correlation-based clustering step, without inverting it; inverse-volatility uses only diagonal volatilities; and 1/N uses no estimated inputs at all. Out-of-sample rankings among these methods are notoriously unstable across universes, samples, and constraints (DeMiguel, Garlappi, and Uppal 2009; Jagannathan and Ma 2003), which is itself an argument for conditioning the choice rather than committing to one method.

Selecting versus diversifying. The most durable challenge to optimization is that it often fails to beat naive diversification: across fourteen models and seven datasets, none consistently outperforms the 1/N rule out of sample, because the estimation window needed to overcome parameter noise is implausibly long (DeMiguel, Garlappi, and Uppal 2009). That result turns 1/N from a straw man into the benchmark any construction method must clear. One response is to combine methods rather than choose among them: blending a sophisticated optimizer with the 1/N target optimizes a bias-variance trade-off (Tu and Zhou 2011), and mixing models can be posed jointly with transaction costs and model uncertainty in large-scale allocations (Hautsch and Voigt 2019), the natural steelman on the diversify side of the present design. Combination and selection are nonetheless distinct: combination averages weights at a point in time, whereas selection picks one method as a function of the state. Whether to combine or to select is itself a classic question in forecasting, where the advantage of simple combination depends on model instability and on the ratio of forecast-error variances (Timmermann 2006). Transposed to construction methods, this is the theoretical antecedent of the disagreement and instability conditions studied here; what it does not supply is an empirical test for

construction methods, nor the observable state that would tell an investor when switching is warranted.

Regimes in portfolio choice. Regime-switching models let expected returns, volatilities, and correlations differ across calm and turbulent states, and conditioning tactical weights on the inferred state has been found to add economic value, though with debated magnitude (Ang and Bekaert 2002; Ang and Timmermann 2012; Kritzman, Page, and Turkington 2012). Regime information is increasingly embedded in the construction algorithm itself (Costa and Kwon 2019; Ciciretti and Bucci 2023), but the common structure is unchanged: the regime reshapes the return distribution, and the investor re-estimates conditional moments and re-optimizes asset weights within a single, fixed framework. The same holds for the factor-timing and tactical-allocation literatures, which condition asset weights rather than the construction method. Schematically, the existing literature studies

$$\text{market state} \rightarrow \text{portfolio weights},$$

whereas the framework here is

$$\begin{aligned} \text{market state} &\rightarrow \text{estimation environment} \rightarrow \text{method reliability} \\ &\rightarrow \text{construction method} \rightarrow \text{portfolio weights}. \end{aligned}$$

The difference is not one of implementation: it changes the object of inference from state-dependent portfolio weights to state-dependent method reliability.

Selecting methods by state. Where the regime does influence the choice of method, it does so through a learned or designed policy rather than a tested hypothesis. The closest work derives, by reinforcement learning, an optimal policy that switches among competing portfolio models under rebalancing costs using macroeconomic information, and shows it can beat the hindsight-best model (Cordoni and Sancetta 2023). A regime-switching decision-support system is a designed policy of the same kind: it maps regime forecasts and an investor's return goal into a discrete risk level, conditional on which it solves a robust mean-variance or a CVaR program, and it is evaluated by backtest performance (Grunnesjö et al. 2026). Both are prescriptive objects; the object here is a set of falsifiable hypotheses about whether the state predicts relative method performance at all. Two questions separate the two programs, and on both the literature is, to our knowledge, silent. First, the standard tool for conditional predictive ability admits either past loss differentials or business-cycle indicators as conditioning variables, but evaluates them in separate test functions rather than testing the incremental content of one over the other (Giacomini and White 2006); we are not aware of a study that tests whether a market-state signal predicts relative model performance beyond momentum across models. Second, we find no study that pits conditional selection against diversification across methods; the benchmarks in this literature are an oracle or single-method baselines.

We make no claim to be the first to condition model choice on the state. The contribution is the specific conjunction: we test whether the state adds information beyond recent performance, compare selection against diversification, make disagreement the boundary condition, tie the value of selection to estimation instability, and use interpretable regimes with placebo controls and a multiplicity correction, which guard against the data snooping that searching over states, methods, and benchmarks invites (White 2000; Hansen 2005). Exhibit 1 summarizes the positioning by the object that the market state conditions.

EXHIBIT 1. Positioning by the object the market state conditions

Literature	Conditions on state?	What changes with the state	Representative work
Regime-switching allocation	Yes	portfolio weights / inputs	Ang and Bekaert (2002); Kritzman, Page, and Turkington (2012); Bae, Kim, and Mulvey (2014)
Regime-aware construction	Yes	constraints within one fixed method	Ciciretti and Bucci (2023)
Regime- and goal-driven decision support	Yes (regime forecast)	risk level and the program solved (robust mean-variance or CVaR)	Grunnesjö et al. (2026)
Model uncertainty / combination	Sometimes	assumptions / combination	Pástor (2000); Anderson and Cheng (2022); Hautsch and Voigt (2019)
Optimal versus naive	No	nothing (unconditional comparison)	DeMiguel, Garlappi, and Uppal (2009); Tu and Zhou (2011)
Combine-vs-select (theory / inference)	Possible; not applied to construction	which model to use	Timmermann (2006); Giacomini and White (2006)
Dynamic model selection	Yes (covariates)	which model (learned policy)	Cordoni and Sancetta (2023)
This paper	Yes (regime)	which construction method, and whether to select at all	—

WHY THE MARKET STATE CAN CARRY INFORMATION

The economic reasoning behind the paper's central claim has three steps: methods differ in how heavily they lean on estimated inputs; the quality of those inputs varies with the market environment; and an observable market state can therefore signal changes in the methods' relative reliability before those changes show up in their recent performance.

Portfolio construction methods occupy a spectrum of estimation dependence, from maximum-Sharpe at the demanding end to equal weighting at the other. A method's reliability, defined formally in the next section, is high when the inputs it relies on are well estimated and degrades as they become noisier. Because methods depend on these inputs to different degrees, they degrade at different rates, and the ranking of methods by reliability reorders as estimation conditions change.

The estimation environment is not constant. The precision of estimated inputs depends on the level and dispersion of volatility, the correlation structure, and the stability of expected returns. These features are not directly observable, but they are correlated with observable market-state variables (realized and implied volatility, credit spreads, and term-structure slope), the same variables used to identify market regimes. The market state is therefore an observable, contemporaneous proxy for the otherwise unobservable estimation environment and carries information about the reordering. The object of conditioning is not the portfolio but the method that produces it.

This is where the distinction from past performance becomes economic rather than statistical. A method's recent relative performance is a lagged, noisy realization of its reliability; the market state is a contemporaneous proxy for the environment that will govern reliability next period. The two coincide only when reliability is highly persistent. They diverge at environment transitions: when the environment shifts, the reliability ranking reorders before recent performance reflects the shift, because recent performance is still dominated by the outgoing environment. These transitions, especially moves into

stress, are also the moments of greatest economic consequence, since that is when estimation-dependent methods do the most damage.

This is why the test must compare information sets rather than test the state in isolation. If the state adds information after conditioning on a strong contemporaneous performance benchmark, the natural explanation is that it proxies the forward-looking estimation environment in a way lagged performance cannot. The claim is not that the state replaces performance information, but that it carries information performance does not.

The same logic locates the boundary condition and the mechanism. Methods disagree most in their exposures precisely when their heterogeneous sensitivities translate the environment into materially different portfolios. Cross-model disagreement is therefore the observable signature of an environment in which the choice of method is consequential, and the marginal value of state information should rise with economically meaningful disagreement. And if the mechanism is estimation stress, the methods the state flags as unreliable should show its symptoms, principally concentrated weights and elevated turnover, providing corroborating, though not causal, evidence for the channel.

HYPOTHESES AND RELIABILITY

Reliability

We define the reliability of a portfolio construction method as its ability to translate its available information into stable, diversified, and economically efficient exposures under the estimation environment it faces. Reliability is reflected in three dimensions: performance (out-of-sample Sharpe, mean, drawdown), stability (weight instability, covariance conditioning), and implementation efficiency (turnover, effective number of assets, concentration). The selection rule deliberately uses performance alone; the stability and implementation dimensions are reserved for ex-post validation of the mechanism, not used as inputs to the selection rule. Reliability is thus the economic concept that motivates why methods can perform differently as the estimation environment changes, not a composite whose weights are tuned to reproduce the performance ranking. The formal metrics are collected in Appendix B.

The information-set hierarchy

For two methods i and k , let $\Delta R_{t+1} = R_{i,t+1} - R_{k,t+1}$ denote the difference in their realized performance over the next holding period, from t to $t + 1$. At date t the investor has two nested information sets: $I_{0,t}$, the methods' past relative performance through t , and $I_{1,t} = I_{0,t} \cup \{S_t\}$, which adds the observable market state S_t . We drop the time subscript when referring to the sets generically or to the selectors built on them. Only the stronger of two relationships is economically informative. The weak relationship,

$$E[\Delta R_{t+1}|S_t] \neq 0,$$

is necessary but not sufficient: it can hold merely because S_t is correlated with recent performance, which is itself predictive through persistence, so a selector exploiting only this could be momentum across methods in disguise (Carhart 1997). The pivotal hypothesis is that the state carries information beyond past performance,

$$E[\Delta R_{t+1}|I_{1,t}] \neq E[\Delta R_{t+1}|I_{0,t}],$$

a statement of conditional predictive ability in the sense of Giacomini and White (2006).

The specification of I_0 is therefore central to the design: a stale or impoverished benchmark would let the state appear informative merely by being more timely. We specify it in two layers (Appendix B): a primary control, the pair's trailing six-rebalance mean performance differential and its most recent value, and a strong benchmark that adds trailing means at thirteen and twenty-six rebalances, two further lags, and a trailing thirteen-rebalance information ratio, so any incremental predictive content of S_t must survive conditioning on a rich history of recent performance rather than merely reflect persistence. Whether this information has economic value is a separate question, addressed by H3.

The hypothesis chain

The framework yields five hypotheses, forming a single logical chain from existence to mechanism.

- **H1 (existence).** Relative method performance is state-dependent; the performance ranking of methods reorders across states.
- **H2 (incremental information, pivotal).** The state adds information beyond past performance: $E[\Delta R_{t+1}|I_{1,t}] \neq E[\Delta R_{t+1}|I_{0,t}]$.
- **H3 (economic value).** The additional information has positive economic value net of transaction costs: the state-conditional selector improves on both performance-chasing and diversification across methods in point estimates out of sample, significantly so against performance-chasing.
- **H4 (boundary condition, secondary).** The incremental value of state-conditional selection is greater when competing methods generate larger economically meaningful differences in exposures. We treat H4 as a secondary, boundary-condition analysis, assessed with a continuous specification rather than a sample split.
- **H5 (mechanism / diagnostic).** The methods the state identifies as less reliable show observable symptoms of estimation instability, particularly concentration and turnover; the covariance condition number is a secondary diagnostic. This is corroborating diagnostic evidence, not a causal identification.

H1 establishes that there is something to exploit; H2 that the state adds information beyond past performance; H3 that this information has economic value after costs; H4 when it pays; and H5 provides evidence on why.

DATA, METHODS, AND DESIGN

Universe and data

The study uses a curated cross-asset universe of twelve exchange-traded funds, each a liquid proxy for a distinct asset class: global equity (SPY, EFA, EEM), the US Treasury curve (IEF, TLT), investment-grade and high-yield credit (LQD, HYG), inflation-linked and international sovereign bonds (TIP, BWX), gold (GLD), broad commodities (DBC), and US real estate (VNQ). Restricting the universe to distinct classes rather than near-duplicate funds ensures that disagreement among construction methods reflects genuine cross-sectional structure. The classes are heterogeneous in risk and return, with annualized volatility ranging from about 6% for inflation-linked and intermediate Treasuries to about 26% for real estate, and maximum drawdowns from -14% to -76% (Exhibit 2). This heterogeneity gives the methods room to disagree; their cumulative paths over the sample are shown in Exhibit A1.

Returns are weekly total returns, with dividends and coupons reinvested, from Bloomberg over October 2007 to September 2026 (990 weekly price dates, 989 weekly return observations). Total-return series matter because price-only data would understate the income-bearing sleeves and bias the comparison against the defensive methods that hold them. Short-dated Treasury bills, represented by the SPDR Bloomberg 1–3 Month T-Bill ETF (BIL), serve as the risk-free rate rather than as an investable asset, and all returns and Sharpe ratios are computed in excess of BIL. BIL is near-cash over the sample (annualized volatility 0.35%, maximum drawdown –0.46%). Excluding it from the risky universe keeps the comparison focused on how methods allocate across genuinely risky assets and avoids the mechanical Sharpe ratio that a minimum-variance rule would earn by concentrating in a near-zero-volatility asset.

EXHIBIT 2. Universe and summary statistics (October 2007 to September 2026)

ETF	Asset class	Ann. return (%)	Ann. volatility (%)	Sharpe	Max drawdown (%)
SPY	US equity	11.85	18.25	0.65	–54.6
EFA	Developed ex-US equity	6.15	19.56	0.31	–59.4
EEM	Emerging-market equity	6.18	23.63	0.26	–62.7
IEF	US Treasury 7–10y	3.15	6.64	0.47	–23.3
TLT	US Treasury 20y+	3.66	14.32	0.26	–47.8
LQD	US investment-grade corporate	4.28	9.05	0.47	–24.7
TIP	US inflation-linked Treasury	3.29	6.11	0.54	–14.2
BWX	International sovereign (ex-US)	0.83	8.88	0.09	–33.5
HYG	US high-yield corporate	5.36	11.15	0.48	–33.0
GLD	Gold	10.42	17.18	0.61	–44.7
DBC	Broad commodities	3.90	19.40	0.20	–75.9
VNQ	US real estate	8.51	25.98	0.33	–69.4
BIL	1–3 months T-Bill	1.32	0.35	3.75	–0.46

NOTES: Weekly total returns over the full sample (990 weekly price dates, 989 weekly return observations); annualized figures. The risk-free rate is proxied by BIL and is not part of the risky universe.

Construction methods

We compare seven long-only construction methods, chosen to span two dimensions, the optimization objective and the portfolio architecture, and through them a broad spectrum of dependence on estimated inputs: maximum-Sharpe, maximum-diversification, minimum-variance, Hierarchical Risk Parity, Hierarchical Equal Risk Contribution, inverse-volatility, and equal weighting (1/N). Ordered from most to least dependent on estimated inputs, they run from maximum-Sharpe, which uses both expected returns and the inverse covariance matrix, to equal weighting, which uses no estimated inputs at all. This ordering is a structural property of how each method uses estimated inputs, not an ex-ante labeling of methods as defensive or aggressive. The covariance matrix is estimated by Ledoit–Wolf linear shrinkage in the main specification. Formal definitions of all seven methods are given in Appendix B.

Regime identification

The market state S_t takes three values (calm, neutral, stress) in the main specification. Consistent with the timing argument above, the detector must be contemporaneous or leading rather than a backward-smoothed filter, since a smoothed state is itself a lagged signal. The main detector assigns states by terciles of the realized volatility of the equally weighted asset portfolio over a thirteen-week trailing window, roughly one quarter, a standard volatility horizon fixed ex ante rather than selected for

significance; thresholds are formed from data available at the decision date only. Realized volatility is the primary state variable because it is the most direct observable proxy for the estimation environment and requires no weighting of components.

Because the thresholds are estimated on an expanding window and volatility is persistent, realized occupancy departs from one-third: of the 221 monthly formation dates, 91 are calm, 87 neutral, and 43 stress; the regime over time is shown in Exhibit A2. A composite stress index (standardized VIX, high-yield spread, dollar index, and sign-reversed curve slope), and a variant adding the MOVE index, are examined as robustness in Appendix C.

Design, costs, and inference

Portfolios are formed every four weeks and held over the subsequent four-week period. All portfolio inputs, including the covariance matrix and any expected-return estimate, use a trailing window of the most recent 104 weeks, strictly prior to the holding period, so no out-of-sample information enters weight formation. At each decision date the procedure estimates inputs, constructs every method's weights, assigns the state, forms the three strategies below, and records gross and net holding-period returns; the window then rolls forward.

Three strategies turn the seven methods into a portfolio decision. The state-conditional selector holds, in each state, the method with the highest mean excess return over all past formation dates in that state, using an expanding history with a minimum of eight observations, below which it falls back to performance-chasing. Performance-chasing, the state-blind benchmark, holds the method with the highest mean excess return over the six most recent rebalances, regardless of state. Diversification across methods holds the equally weighted average of the seven methods' portfolios. The selection rules are specified in Appendix B.

All comparisons are reported net of transaction costs; because a method-switching selector is inherently high-turnover, the gross-to-net gap is reported explicitly rather than deferred (Appendix A), and no turnover penalty enters the selection rule. Sharpe-ratio differences are tested with a studentized circular block bootstrap (Ledoit and Wolf 2008), and the model confidence set (Hansen, Lunde, and Nason 2011) identifies the methods statistically indistinguishable from the best; inference details, placebos, and the multiplicity correction are in Appendices C and D, and the full walk-forward protocol in Appendix A.

EMPIRICAL RESULTS

The performance ranking reorders across states (H1)

The relative performance of the construction methods depends strongly on the market state. Exhibit 3 reports each method's annualized Sharpe ratio and mean excess return within each state, and Exhibit 4 shows the same ranking graphically. The reordering is economically large and, for the sharpest pair, highly significant. Maximum-Sharpe, the return-seeking method, swings from a Sharpe of 1.12 in calm markets to -0.35 in stress, while the estimation-free equal-weight portfolio moves in the opposite direction, from 0.20 in calm to 0.79 in stress. The pair whose ranking reverses most is equal-weighting versus maximum-Sharpe: their Sharpe gap swings from clearly favoring maximum-Sharpe in calm markets to clearly favoring equal-weighting in stress (reversal statistic -2.06 , bootstrap $p < 0.001$). This is the pattern the estimation-dependence spectrum

anticipates: the methods that rely most heavily on estimated inputs are the ones whose relative performance is most state-sensitive.

Two features matter for the interpretation that follows. First, the reversal is visible in both mean return and the Sharpe ratio but is sharper in the latter: the return-seeking method leads in calm and neutral markets yet delivers a negative excess return in stress, and the risk-adjusted reversal is larger still. Second, the reordering is far more pronounced under the realized-volatility detector than under the broader composite detector, foreshadowing a result made explicit below: the predictive content of the state is specific to the volatility detector..

EXHIBIT 3. Method performance across market states

Panel A. Annualized Sharpe ratio, by method and state

Method	Calm	Neutral	Stress
Maximum Sharpe	1.12	1.17	−0.35
Maximum diversification	0.84	0.80	0.59
Minimum variance	0.54	0.45	0.51
Hierarchical Risk Parity	0.34	0.45	0.57
Hierarchical Equal Risk Contribution	0.27	0.21	0.61
Inverse volatility	0.33	0.69	0.75
Equal weighting (1/N)	0.20	0.89	0.79

Panel B. Annualized mean excess return (%), by method and state

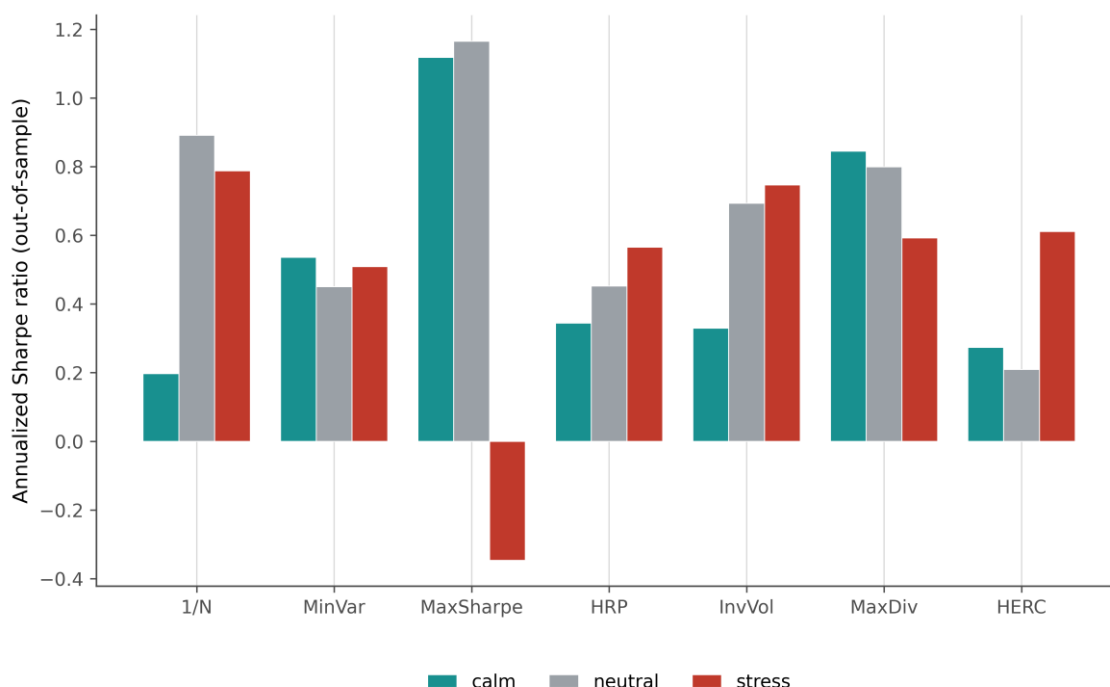
Method	Calm	Neutral	Stress
Maximum Sharpe	7.0	10.0	−4.0
Maximum diversification	4.0	5.0	5.0
Minimum variance	2.0	2.0	4.0
Hierarchical Risk Parity	2.0	3.0	5.0
Hierarchical Equal Risk Contribution	1.0	1.0	5.0
Inverse volatility	2.0	4.0	8.0
Equal weighting (1/N)	1.0	6.0	10.0

Panel C. Reversal test

Most-reversing pair	Equal weighting vs Maximum Sharpe
Sharpe gap, calm state	−0.92
Sharpe gap, stress state	+1.13
Reversal statistic	−2.06
Bootstrap p-value	< 0.001

NOTES: Panels A and B: annualized out-of-sample Sharpe ratio and mean excess return for each method by market state (trailing-realized-volatility terciles). Panel C: the pair whose Sharpe ranking reverses most between calm and stress, with a block-bootstrap reversal test.

EXHIBIT 4. Performance by market state



The state predicts beyond recent performance (H2)

This is the pivotal test: H1 shows that performance differs across states; H2 asks whether the state predicts those differences once recent relative performance is taken into account. Exhibit 5 reports the conditional-predictive-ability test for every pair of methods: the forward performance differential is regressed on a contemporaneous performance-only control and on the state, and the state coefficients are jointly tested. The state carries significant incremental information in eight of twenty-one pairs, and the significance is concentrated in the most estimation-dependent method: every pair involving maximum-Sharpe rejects the null (Wald up to 17.9, $p < 0.001$), with the state predicting that the return-seeking method underperforms in stress. This is not a mechanical restatement of the reversal in Exhibit 3, because the test conditions on recent relative performance; the state adds information over and above momentum in relative method performance, and all six maximum-Sharpe pairs survive a Holm–Bonferroni multiplicity correction. This concentration is what the mechanism predicts *ex ante* rather than a limitation, and the signal is not confined to maximum-Sharpe: the two remaining significant pairs, which also survive the correction, set maximum-diversification, another method that relies on the full covariance matrix, against the estimation-light 1/N and inverse-volatility rules, with the estimation-dependent method again losing ground in stress.

The result is not spurious. Three placebo checks confirm it (Appendix C): a permutation placebo that shuffles the state assignment collapses the statistic toward its no-information distribution; a wrong-mapping placebo that routes each state to the worst rather than the best method within it is beaten by the correct mapping out of sample ($\Delta\text{Sharpe } 0.78$, $p < 0.001$); and a strong-benchmark test that expands the performance-only control to seven trailing statistics leaves the state's information intact, evidence that it is not recent relative performance in disguise. The signal is also specific to the volatility

detector: broader composite indices that dilute the volatility signal with the dollar and the curve do not reproduce it (Appendix C).

Most important, the predictive content is not an artifact of sample-mean instability in a single method. Because maximum-Sharpe is the only method that uses estimated expected returns, its state-dependence could reflect the noise in the sample mean rather than a broader property of estimation-dependent construction. Exhibit 6 re-estimates maximum-Sharpe with Bayes–Stein and James–Stein expected returns, holding the covariance and optimizer fixed. The performance reversal is preserved; shrinkage weakens the signal under Bayes–Stein (mean Wald 13.5 to 8.4, five of six pairs significant and the sixth at $p = 0.050$) but does not remove it, and James–Stein leaves it essentially intact. The Bayes–Stein variant remains a genuinely return-seeking portfolio rather than minimum-variance in disguise, tracking sample maximum-Sharpe (return correlation 0.95) far more closely than minimum-variance (0.63) and showing the same rise in concentration and turnover in stress, while maximum-diversification, which uses the covariance but no estimated mean, does not reverse across states (Appendix C). The residual state information is therefore consistent with a return-seeking objective under noisy estimation rather than with covariance conditioning alone.

A leave-one-episode-out exercise (Appendix C) shows the result is robust to removing any single calendar year: all six maximum-Sharpe pairs remain significant in every case, including when 2022, the rate shock, is excluded, so the finding does not rest on one episode.

EXHIBIT 5. Giacomini–White conditional-predictive-ability test, all method pairs

Pair	Wald	p	calm coef	calm se	stress coef	stress se	sig 5%
1/N-MinVar	5.59	0.061	-0.0032	0.0018	+0.0016	0.0026	
1/N-MaxSharpe	17.89	< 0.001	-0.0008	0.0027	+0.0144	0.0038	*
1/N-HRP	4.97	0.083	-0.0023	0.0014	+0.0014	0.0021	
1/N-InvVol	5.02	0.081	-0.0014	0.0007	+0.0001	0.0011	
1/N-MaxDiv	15.13	0.001	-0.0032	0.0018	+0.0029	0.0024	*
1/N-HERC	3.86	0.145	-0.0036	0.0021	-0.0002	0.0030	
MinVar-MaxSharpe	12.74	0.002	+0.0023	0.0028	+0.0134	0.0038	*
MinVar-HRP	2.30	0.316	+0.0008	0.0007	-0.0006	0.0011	
MinVar-InvVol	5.08	0.079	+0.0017	0.0011	-0.0018	0.0017	
MinVar-MaxDiv	0.72	0.696	+0.0002	0.0010	+0.0012	0.0014	
MinVar-HERC	1.16	0.561	+0.0002	0.0013	-0.0013	0.0014	
MaxSharpe-HRP	11.99	0.002	-0.0014	0.0026	-0.0131	0.0038	*
MaxSharpe-InvVol	15.43	< 0.001	-0.0006	0.0025	-0.0141	0.0038	*
MaxSharpe-MaxDiv	11.30	0.004	-0.0021	0.0026	-0.0119	0.0036	*
MaxSharpe-HERC	11.74	0.003	-0.0025	0.0029	-0.0149	0.0044	*
HRP-InvVol	4.85	0.088	+0.0009	0.0007	-0.0013	0.0011	
HRP-MaxDiv	3.27	0.195	-0.0005	0.0012	+0.0016	0.0012	
HRP-HERC	0.27	0.874	-0.0005	0.0012	-0.0005	0.0012	
InvVol-MaxDiv	15.35	< 0.001	-0.0016	0.0014	+0.0030	0.0016	*
InvVol-HERC	1.87	0.393	-0.0019	0.0015	+0.0002	0.0019	
MaxDiv-HERC	2.43	0.296	-0.0004	0.0014	-0.0026	0.0017	

NOTES: Each test uses 215 observations (221 monthly formation dates less a six-rebalance burn-in for the trailing performance control). For each method pair, the next-period performance differential is regressed on a performance-only control and the market state; state coefficients ("calm coef", "stress coef") are relative to neutral and jointly tested, with HAC standard errors.

EXHIBIT 6. Expected-return robustness: state predictability survives stabilizing the mean

Panel A. MaxSharpe reliability under alternative expected-return estimators (annualized Sharpe by state).

μ estimator	Calm	Neutral	Stress
Sample mean	1.12	1.17	−0.35
Bayes–Stein	1.10	1.12	−0.14
James–Stein	1.10	1.17	−0.35

Panel B. Incremental state information (Giacomini–White Wald, p in parentheses), pairs involving MaxSharpe.

Pair	Sample	Bayes–Stein	James–Stein
1/N – MaxSharpe	17.9 (<0.001)	12.2 (0.002)	18.9 (<0.001)
MinVar – MaxSharpe	12.7 (0.002)	7.5 (0.023)	14.8 (0.001)
HRP – MaxSharpe	12.0 (0.002)	7.3 (0.026)	14.2 (0.001)
InvVol – MaxSharpe	15.4 (<0.001)	10.1 (0.006)	17.0 (<0.001)
MaxDiv – MaxSharpe	11.3 (0.004)	6.0 (0.050)	13.9 (0.001)
HERC – MaxSharpe	11.7 (0.003)	7.4 (0.025)	14.3 (0.001)
Pairs significant (5%)	6 / 6	5 / 6	6 / 6
Mean Wald	13.5	8.4	15.5

Panel C. The Bayes–Stein variant is a return-seeking portfolio, not minimum-variance.

Diagnostic	Value
Weight distance $\frac{1}{2}\ \cdot\ _1$: Bayes–Stein vs sample MaxSharpe	0.22
Weight distance $\frac{1}{2}\ \cdot\ _1$: Bayes–Stein vs MinVar	0.50
Return correlation : Bayes–Stein vs sample MaxSharpe	0.95
Return correlation : Bayes–Stein vs MinVar	0.63
Bayes–Stein concentration ($1/N_{\text{eff}}$), calm $\rightarrow$ stress	0.27 $\rightarrow$ 0.46
Bayes–Stein turnover, calm $\rightarrow$ stress	0.11 $\rightarrow$ 0.18
MaxDiv Sharpe ($\mu \propto \sigma$, no sample mean): calm / neutral / stress	0.84 / 0.80 / 0.59

NOTES: Maximum–Sharpe is re-estimated with each expected-return estimator; the covariance (Ledoit–Wolf), optimizer, sample, and state assignment are unchanged across columns. Panel B: Wald statistics on the state coefficients, as in Exhibit 5. Panel C: weight distances averaged over formation dates; concentration is $1/N_{\text{eff}}$; turnover is one-way per rebalance.

Economic value net of costs (H3)

H2 establishes that the state contains incremental predictive information; H3 asks whether exploiting that information is economically useful after transaction costs, comparing the state-conditional selector with performance-chasing and with diversification across methods. The selector is stable rather than noise-driven: it learns on an expanding history, and outside a two-year warm-up the selected method changes at fewer than 3% of formation dates in calm markets and never in stress, its modal choices being maximum–Sharpe in calm markets and equal weighting in stress (Appendix C).

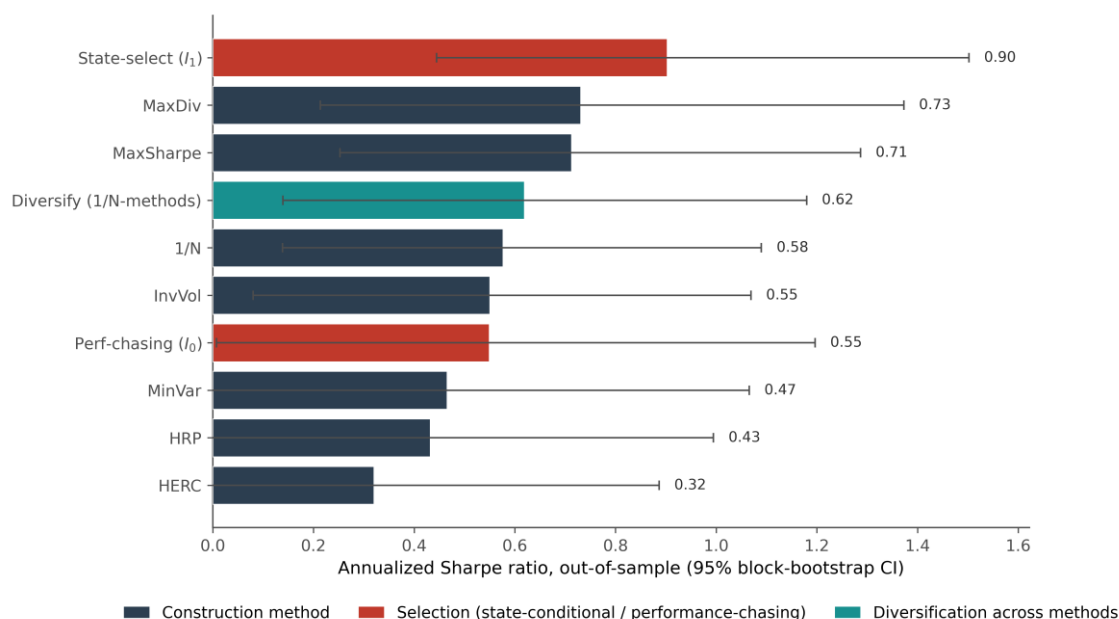
Exhibit 7 reports the out-of-sample Sharpe ratios of all comparators with bootstrap confidence intervals and the model confidence set. Gross of costs, the state-conditional selector has the highest point estimate, 0.90 against 0.73 for maximum-diversification, the next highest. This does not imply a statistically isolated winner: the model confidence set retains several methods, including the state-conditional selector, diversification, and maximum-Sharpe, as statistically indistinguishable from the best. Exhibit 8 shows the cumulative out-of-sample paths of the three rules: the state-conditional selector pulls steadily ahead of both performance-chasing and diversification over the sample.

The economic result differs across the two benchmarks. Net of transaction costs, the state-conditional selector significantly outperforms performance-chasing (net Sharpe-ratio difference 0.38, bootstrap $p = 0.03$; Exhibit 9), because the performance-chasing rule turns the portfolio over far more (350% versus 261% annualized) and pays for it: its net Sharpe ratio falls to 0.46 against the selector's 0.83. The two rules differ in memory as well as in state-conditioning, so we decompose the difference with a third selector that shares the state-conditional selector's expanding memory but ignores the state. It earns a net Sharpe ratio of 0.58, no better than diversification (0.59): longer memory alone adds 0.12 over performance-chasing, and conditioning on the state adds a further 0.25. Neither increment is individually significant ($p = 0.47$ and 0.13), so the split is descriptive rather than inferential: about two-thirds of the difference is associated with state conditioning, and the predictive tests show that the state's information is not explained by recent performance at horizons up to twenty-six rebalances. The selector's advantage over diversification across methods, by contrast, is positive but does not reach conventional significance ($p \approx 0.11$): when the methods largely agree, diversification is an efficient default. The state-conditional selector therefore delivers a demonstrable economic gain relative to performance-chasing, while selection versus diversification remains a closer, disagreement-dependent comparison that the next section takes up.

The economic result is more sensitive to the detector window than the predictive one (Appendix C). All six maximum-Sharpe pairs remain significant at thirteen, seventeen, and twenty-six weeks, but only two of six at the shortest window, nine weeks; the gross Sharpe difference over performance-chasing is significant only at thirteen and seventeen weeks. We therefore read the economic magnitude as indicative rather than precisely pinned down, while the predictive result holds at windows of a quarter or longer.

EXHIBIT 7. Out-of-sample performance and Sharpe-difference tests

Panel A. Annualized Sharpe ratios with 95% bootstrap confidence intervals



Panel B. Sharpe-difference tests (Ledoit–Wolf)

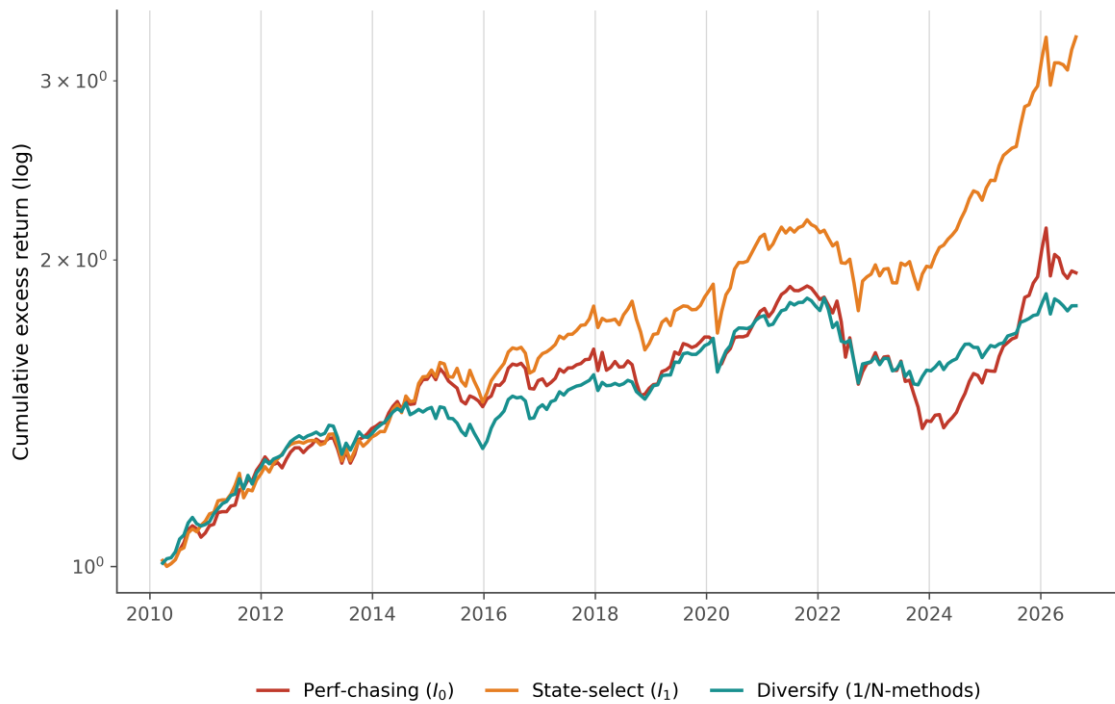
Comparison	Δ SR	HAC p	Bootstrap p
Diversify (1/N-methods) vs Perf-chasing (I_0)	0.07	0.666	0.669
Diversify (1/N-methods) vs State-select (I_1)	−0.28	0.057	0.055
State-select (I_1) vs Perf-chasing (I_0)	0.35	0.030	0.029
Diversify (1/N-methods) vs MaxSharpe	−0.09	0.624	0.611
Diversify (1/N-methods) vs MinVar	0.15	0.061	0.073
MaxDiv vs State-select (I_1)	−0.17	0.371	0.399

Panel C. Model confidence set

	in 90% MCS	Ann. Sharpe
State-select (I_1)	yes	0.90
MaxDiv	yes	0.73
MaxSharpe	yes	0.71
Diversify (1/N-methods)	yes	0.62
1/N	yes	0.58
InvVol	yes	0.55
Perf-chasing (I_0)	yes	0.55
MinVar	no	0.47
HRP	no	0.43
HERC	no	0.32

NOTES: Panel A: each strategy's annualized out-of-sample Sharpe ratio with a 95% bootstrap confidence interval. Panel B: pairwise annualized Sharpe-ratio differences (Δ SR) with HAC and studentized-bootstrap p-values. Panel C: 90% model confidence set; "yes" denotes membership. I_1 is the state-conditional selector, I_0 performance-chasing. All figures are gross of transaction costs; net figures in Exhibit 9.

EXHIBIT 8. Selection versus diversification: out-of-sample cumulative performance



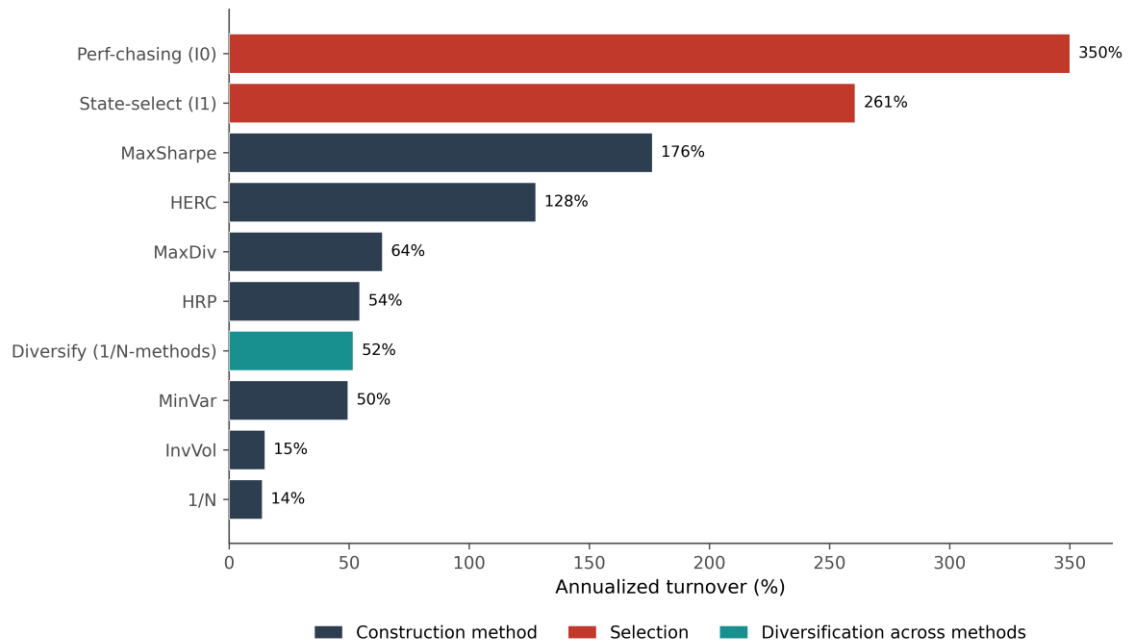
NOTES: Cumulative out-of-sample excess return (log scale) of the state-conditional selector (I_1), performance-chasing (I_0), and diversification across methods, October 2010 to September 2026. All series are net of transaction costs and net of the risk-free rate.

EXHIBIT 9. Turnover and net-of-cost Sharpe by strategy

Panel A. Gross and net Sharpe ratios and turnover, by strategy

	Sharpe gross	Sharpe net	Ann. turnover %
State-select (I_1)	0.89	0.83	260.7
MaxDiv	0.73	0.70	63.9
MaxSharpe	0.71	0.67	176.3
Diversify (1/N-methods)	0.61	0.59	51.7
1/N	0.57	0.57	13.9
Perf-chasing (I_0)	0.55	0.46	350.1
InvVol	0.54	0.54	15.0
MinVar	0.46	0.44	49.5
HRP	0.42	0.41	54.4
HERC	0.32	0.28	127.7

Panel B. Annualized turnover by strategy



Panel C. Net-of-cost Sharpe-difference tests

Detector	Comparison	Sharpe (I ₁), net	ΔSharpe, net	Bootstrap p
vol	I ₁ vs I ₀	0.83	0.38	0.026
vol	I ₁ vs Diversify	0.83	0.24	0.113
vol	I ₁ vs Expanding, no state	0.83	0.25	0.126
composite	I ₁ vs I ₀	0.49	0.03	0.858
composite	I ₁ vs Diversify	0.49	-0.11	0.475

NOTES: All Sharpe ratios are net of transaction costs (ten basis points per unit of turnover; Appendix A); turnover is annualized one-way. Panel A: gross and net Sharpe ratios and turnover by strategy. Panel B: annualized turnover. Panel C: net-of-cost Sharpe differences, state-conditional selector versus each benchmark, with bootstrap p-values; the expanding benchmark shares the selector's memory but ignores the state (Appendix C). Gross figures and the model confidence set: Exhibit 7.

When selection pays: the disagreement boundary (H4)

H3 establishes that state-conditional selection can create economic value; H4 asks when that value is greatest. The mechanism predicts that the value of selecting among methods should rise with their disagreement, because that is when the choice is consequential. Exhibit 10 tests this boundary condition continuously, regressing the selector's forward edge over each benchmark on cross-model disagreement while controlling for the market state. Disagreement enters as average pairwise weight distance and as average pairwise ex-ante tracking error, the latter a direct measure of economically meaningful differences in exposures. The evidence is directional but weaker than a confirmatory test would require. The edge increases with disagreement in the predicted direction in every specification. For the non-mechanical contrast, selection versus performance-chasing, where both rules pick a single method, the slope is significant under the tracking-error measure (one-sided $p = 0.02$) and marginal under weight distance ($p = 0.08$). The advantage over diversification is significant under weight distance but not

under tracking error, and this contrast is in any case partly mechanical, since a $1/N$ average of more divergent portfolios necessarily lies further from any single method. We therefore read H4 as boundary-condition evidence: the case for selection over performance-chasing strengthens when methods diverge, consistent with the mechanism, but the evidence is suggestive rather than conclusive.

EXHIBIT 10. The selection edge and cross-model disagreement (continuous specification)

Selection edge	Weight disagreement	Ex-ante tracking error
vs performance-chasing ($I_1 - I_0$)	+1.82 (0.075)	+3.23 (0.018)
vs diversification ($I_1 - \text{Div}$) †	+2.64 (0.004)	+0.87 (0.230)

NOTES: $\text{edge}_{t+1} = a + bD_t + (\text{state dummies}) + \varepsilon_{t+1}$, estimated by OLS with Newey–West standard errors. The coefficient b is the annualized change in the selection edge (percentage points) per one-standard-deviation increase in disagreement D_t ; one-sided p (H4: $b > 0$) in parentheses. Disagreement: average pairwise weight distance and average pairwise ex-ante tracking error (Appendix B). † Significant under the weight measure only; see text.

The estimation-stress channel (H5)

H4 asks when state-conditional selection is most valuable; H5 asks whether the methods the state flags as unreliable display the symptoms expected under an estimation-stress channel. The evidence is consistent with the channel, though partial. The clearest symptom is concentration: the most estimation-dependent method, which the state predicts to fail in stress, becomes markedly more concentrated precisely in the stressed state ($1/N_{\text{eff}}$ rising from 0.35 in calm to 0.48 in stress; Exhibit D1), while the concentration of the other methods changes far less, consistent with unstable, estimation-driven weights. Turnover is less discriminating: it rises in stress for maximum-Sharpe, but proportionally more for the estimation-light methods, so it reflects the stressed environment rather than the method. The covariance condition number does not rise in stress (68.0 in calm, 63.0 in stress), consistent with the Ledoit–Wolf shrinkage used in the main specification, which stabilizes covariance conditioning across states. The mechanism evidence therefore rests on concentration. As stated in H5, this is diagnostic support for the channel rather than a causal identification of it.

The cross-sectional pattern is consistent with the same channel. In equities-only and fixed-income-only universes the state carries little predictive information (none and two of twenty-one pairs significant, respectively, none involving maximum-Sharpe); in the multi-asset universe, where heterogeneous asset classes give the methods room to diverge, the predictive content is concentrated (Appendix C).

USING THE FRAMEWORK IN PRACTICE

For a portfolio manager, the practical value of these results is that they turn the choice of construction method from a one-time modeling commitment into an ongoing, state-conditional judgment, using information already available in the investment process. The framework requires no new construction method and no complex regime detector: the prevailing level of realized volatility, a single transparent variable, carries information about when the most estimation-dependent method is becoming unreliable, beyond what that method's own recent performance reveals.

The information is best used defensively. The state is most informative about the periods in which estimation-dependent methods, maximum-Sharpe foremost, deteriorate:

their risk-adjusted performance is strongest in calm markets and weakest in stress, when they also become more concentrated. Because the state carries information beyond trailing performance, it can flag an estimation-dependent method as becoming unreliable before the deterioration is fully reflected in its lagged-performance signal. In practice this argues less for picking a winner in good times than for leaning away from a fragile method as the environment turns, toward the estimation-light methods that hold up when inputs become noisy.

The framework also indicates when the choice of method is worth making at all. When the candidate methods produce similar portfolios, the choice among them scarcely matters and simple diversification across methods is an efficient default. When they disagree, which is itself observable in real time from the dispersion of their weights, the case for conditioning on the state is stronger. Cross-model disagreement can therefore serve as a rough, real-time guide to when method selection is worth the effort, though on the present evidence it is a guide rather than a decisive trigger, and the advantage of selection over diversification is suggestive rather than established.

Two cautions frame the use of the signal. First, it is a complement to existing risk budgeting, not a replacement: a manager can run it alongside a volatility- or risk-parity-based process and read the state as an additional input on which construction method to trust, rather than as a standalone allocation rule. Second, the economic case is asymmetric. The gain over performance-chasing is significant net of transaction costs, though its magnitude is sensitive to the detector window; the gain over diversification is positive but imprecise. The finding should be used as a reliability signal with a demonstrable payoff against performance-chasing, not as a turnkey source of alpha.

CONCLUSION

The choice of how to build a portfolio is usually treated either as a fixed methodological commitment or as a source of uncertainty to be diversified away. This paper shows that it is neither. The relative performance of construction methods is state-dependent, and the market state carries usable, observable information about which method to trust. The signal is not exotic: the prevailing level of realized volatility predicts when the most estimation-dependent method becomes unreliable, beyond what the method's own recent performance reveals. This incremental predictive content survives placebo tests, a multiplicity correction, and shrinkage of the expected-return input, which weakens but does not remove it.

The evidence forms a clear hierarchy. The performance ranking reorders sharply across states; the state adds incremental information about that reordering, concentrated in the method most exposed to estimation error; exploiting it is significantly valuable against performance-chasing, though positive and imprecise against diversification; and the flagged method shows the concentration symptom of estimation stress. The predictive result provides the strongest evidence and is consistent with the estimation-stress channel; the economic and mechanism results point in the same direction without, at this sample length, being individually decisive.

These conclusions are bounded by the setting. The evidence comes from a single multi-asset universe of liquid ETFs over nineteen years, with a realized-volatility detector and just over two hundred monthly decisions; the economic gain over diversification is positive but not statistically established, the magnitude of the gain over performance-chasing is sensitive to the detector window, and the mechanism evidence is diagnostic

rather than causal. Larger and less liquid universes, alternative state variables, and settings with liabilities are natural next tests, as is eliciting when a manager should act on cross-model disagreement rather than diversify.

The broader lesson for the investment process is that model uncertainty at the construction stage is partly predictable and therefore partly manageable. Rather than committing to a single construction method, or assuming that diversification across methods is always the best response, a manager can monitor the market state, most simply the volatility regime, for information about when an estimation-dependent method is becoming less reliable. The choice of construction method is not a settled question to be answered once; it is a state-dependent decision, and the state contains information the manager can use.

APPENDIX A. DATA AND IMPLEMENTATION

A.1 Estimation and formation

Portfolios are formed every four weeks (approximately monthly) and held over the subsequent four-week period. All estimation, of covariances and of any expected-return inputs, uses a trailing window of the most recent 104 weeks (two years), strictly prior to the holding period, so that no out-of-sample information enters weight formation. Throughout, $N = 12$ denotes the number of risky assets and r_t the vector of period- t weekly excess returns; the walk-forward protocol is detailed in Section A.3.

A.2 Regime-detector inputs

The robustness detectors use five standard, observable stress indicators, also weekly and from Bloomberg over the same sample: the CBOE volatility index (VIX), the option-adjusted spread of the US high-yield corporate bond market (HY OAS), the ICE US dollar index (DXY), the slope of the Treasury curve (ten-year minus three-month yield), and the ICE BofA MOVE index of Treasury-rate implied volatility. The first four define the composite stress index; MOVE, which captures rates volatility as distinct from equity volatility, enters only its extended variant (Appendix C). The primary detector uses none of them: it is the realized volatility of the equally weighted asset portfolio. Over the sample the VIX ranges from 9 to 79, the high-yield spread from 2.5% to 20%, the curve slope from -1.8% (inverted) to $+3.8\%$ (steep), and MOVE from 37 to 265, so the indicators register the full span of calm and stressed conditions. These series are not part of the investable universe; they enter only as conditioning information.

A.3 Transaction costs

Let c denote the per-unit-of-turnover cost. The net holding-period return of a method is its gross return less c times its turnover, and the net return of the selector additionally bears the cost of switching methods at state changes. All headline comparisons for H3 are reported net of costs under a main assumption of ten basis points per unit of turnover, charged on both sides of each trade, with robustness reported at five and twenty basis points. Because a selector that switches methods across regimes is inherently high-turnover, the gross-to-net gap is reported explicitly rather than deferred to a robustness appendix; it is the single result most likely to overturn an apparent gross gain.

A.4 Walk-forward protocol

At each monthly decision date t , using only information through t : estimate μ and Σ on the trailing window; construct the weight vector for every method; assign the state S_t ; form the state-conditional selection and every benchmark; and record gross and net holding-period returns, the reliability metrics, and the disagreement measures D_t . The window then rolls forward one month and the process repeats. This produces, for every method and every benchmark, an out-of-sample return series tagged by the decision-date state, from which all tests are computed.

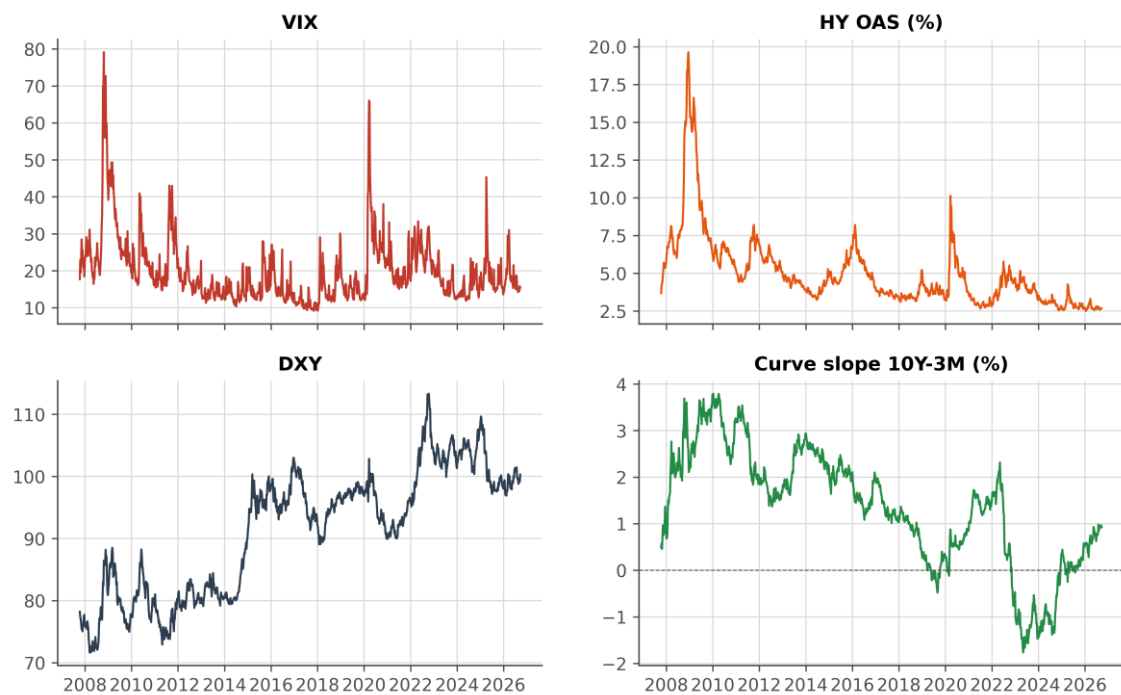
A.5 Exhibits

EXHIBIT A1. ETF cumulative total return, base 100

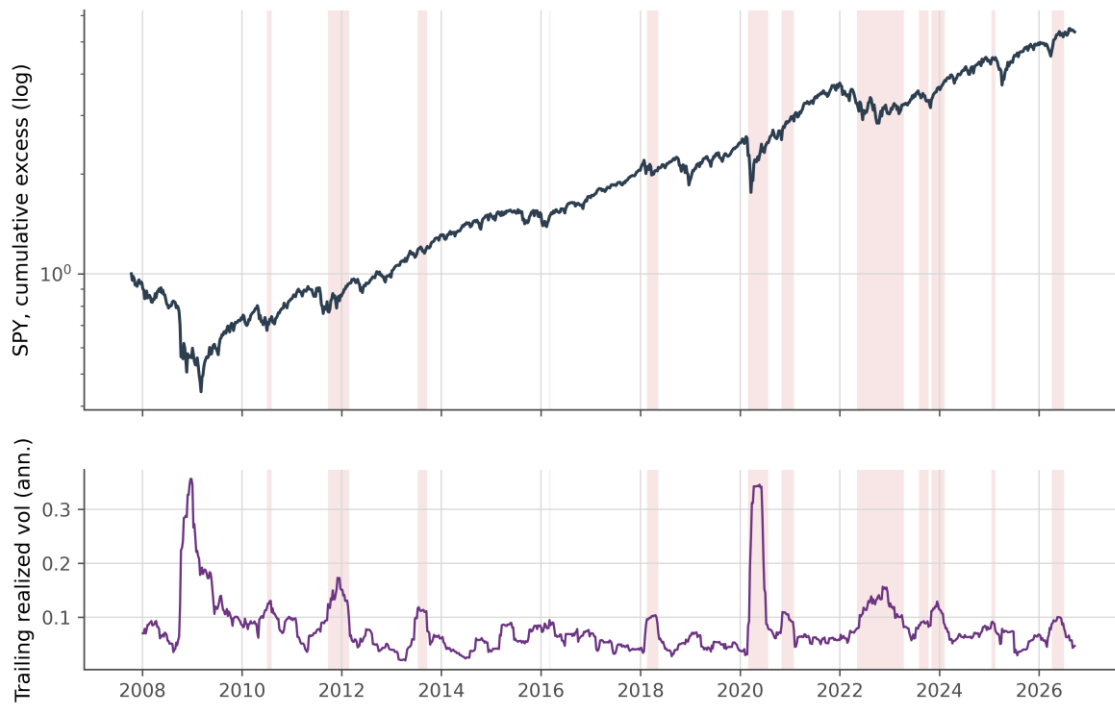


EXHIBIT A2. Market environment and the volatility regime over time

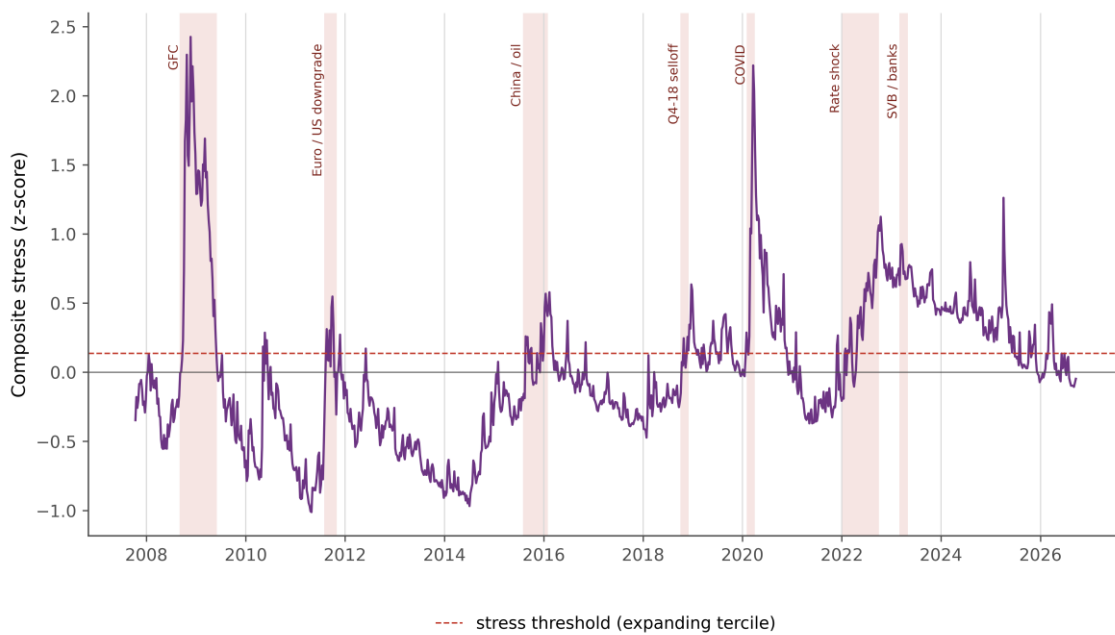
Panel A. Market stress indicators (composite detector inputs).



Panel B. Realized-volatility regime over time.



Panel C. Composite stress index vs known crisis episodes (robustness detector)



NOTES: Panel A: the four inputs to the composite detector (Appendix A.2). Panel B: the realized-volatility regime used as the main detector. Panel C: the composite stress index against known crisis episodes (robustness only; Appendix C).

APPENDIX B. CONSTRUCTION METHODS AND THE SELECTION RULE

B.1 Candidate construction methods

We consider seven long-only construction methods, chosen to span two conceptually distinct dimensions, the optimization objective and the portfolio architecture, and through them the full spectrum of dependence on estimated inputs. Let Σ denote the estimated covariance matrix of the N asset returns over the trailing window, σ the vector of estimated volatilities, and μ the vector of estimated mean excess returns. Every method produces fully invested, long-only weights ($w' \mathbf{1} = 1, w \geq 0$). Ordered from most to least dependent on estimated inputs, the methods are:

1. Maximum Sharpe (tangency):

$$w = \operatorname{argmax}_w \frac{w' \mu}{\sqrt{w' \Sigma w}}$$

It uses both expected returns and the inverse covariance matrix and is therefore the most estimation-dependent method among those considered. When the maximum-ratio program is infeasible over a window, as happens when no asset's estimated excess return exceeds the risk-free rate, we fall back to minimum variance.

2. Maximum diversification. It maximizes the diversification ratio, using the volatilities and the full covariance matrix.

$$w = \operatorname{argmax}_w \frac{w' \sigma}{\sqrt{w' \Sigma w}}$$

3. Minimum variance. It discards expected returns but its solution still depends on the full covariance matrix, so it remains exposed to covariance estimation error and ill-conditioning.

$$w = \operatorname{argmin}_w w' \Sigma w$$

4. Hierarchical Risk Parity (HRP; López de Prado 2016). Assets are clustered by a correlation-distance metric and weights are set by recursive bisection down the resulting tree, allocating inversely to cluster variance. It replaces matrix inversion with a clustering step, trading theoretical optimality for robustness to covariance estimation error.

5. Hierarchical Equal Risk Contribution (HERC; Raffinot 2018b). Assets are clustered as in HRP, but weight is allocated across clusters by equalizing their risk contributions, and within each cluster by naive risk parity, rather than by inverse-variance bisection.

6. Inverse volatility (naive risk parity). Weights are normalized to sum to one. It uses only the diagonal of the covariance matrix, avoiding both expected returns and matrix inversion; it coincides with equal-risk-contribution weighting when all pairwise correlations are equal.

$$w_j \propto \frac{1}{\sigma_j}$$

7. Equal weighting (1/N). It uses no estimated inputs and is, by construction, immune to estimation error in the inputs used to construct the weights, serving as the estimation-free benchmark.

$$w_j = \frac{1}{N}$$

These span the objective dimension (return/risk maximization, variance minimization, risk balancing, naive allocation) and the architecture dimension (flat optimization versus hierarchical clustering), ordering along a single estimation-dependence spectrum from maximum-Sharpe, most exposed, to equal weighting, immune. The covariance matrix is estimated by Ledoit–Wolf linear shrinkage (Ledoit and Wolf 2004) in the main specification; the sample covariance is used in the robustness analysis.

B.2 The information-set test for H2

The pivotal hypothesis is tested as conditional predictive ability (Giacomini and White 2006). For an ordered pair of methods (i, k) , let $\Delta R_{t+1} = R_{i,t+1} - R_{k,t+1}$ denote their realized performance differential over the next holding period. Let h_t be a vector of test functions measurable with respect to the performance-only information set $I_{0,t}$, and let s_t collect the market-state indicators

$$s_t = (\mathbf{1}\{S_t = 1\}, \dots, \mathbf{1}\{S_t = K - 1\})'.$$

We estimate the predictive regression

$$\Delta R_{t+1} = \alpha + \beta' h_t + \gamma' s_t + \varepsilon_{t+1}, \quad (B.1)$$

and test $H_0 : \gamma = 0$ against the alternative that the state adds predictive content beyond past performance. The test statistic is a heteroskedasticity-and-autocorrelation-robust Wald statistic on γ , asymptotically χ^2 with $K - 1$ degrees of freedom under the null. Rejection is the evidence for H2.

Specification of h . In the *primary* specification h comprises two performance-only controls: the trailing six-rebalance mean of the pair's performance differential and its one-period lag. The *strong benchmark* expands h to seven controls: trailing means of the differential over six, thirteen, and twenty-six rebalances; its first three lags; and a trailing thirteen-rebalance information ratio (the mean differential divided by its standard deviation). Because h contains only functions of past relative performance and no volatility term, the state's information is genuinely incremental. Both h and s are frozen before estimation.

B.3 The state-conditional selection rule

The state-conditional selector maps each state to the method with the highest mean realized performance among all past formation dates classified in that state, using an *expanding* history that contains only information available at the decision date: for each state it holds, over the following month, the weights of the method that has performed best in that state. To guard against learning a winner from too few observations, the rule requires at least eight past formation dates in the current state and otherwise falls back to performance-chasing; this fallback binds only during the first two years of the out-of-sample period (24 of 221 formation dates, the last in August 2013). Performance-chasing, the state-blind comparator, selects the method with the best trailing six-rebalance mean return and therefore uses I_0 only. By construction, neither selector optimizes over transaction costs, so the incremental value attributed to the state cannot reflect a tuned cost penalty.

B.4 Cross-model disagreement

H4 requires a measure of how much the methods disagree. The primary measure is *weight disagreement*, the mean pairwise weight distance across method pairs (i, k) ,

$$D_t^W = \underset{(i,k)}{\text{mean}} \frac{1}{2} \|w_{i,t} - w_{k,t}\|_1, \quad (B.2)$$

largest exactly when the methods' heterogeneous sensitivities translate the estimated inputs into materially different portfolios. We complement it with an economically weighted counterpart, the mean pairwise *ex-ante tracking error*,

$$D_t^R = \underset{(i,k)}{\text{mean}} \sqrt{(w_{i,t} - w_{k,t})' \Sigma_t (w_{i,t} - w_{k,t})}, \quad (B.3)$$

which scales weight differences by the estimated return covariance, so disagreement in high-covariance directions counts more than disagreement in negligible ones. H4 is assessed by regressing the forward edge of the state-conditional selector over two benchmarks (performance-chasing and diversification across methods) on disagreement entered continuously, controlling for the market state. The advantage over diversification is partly mechanical, because a 1/N average of more divergent portfolios necessarily lies further from any single method; the comparison against performance-chasing, where both rules hold a single method, is therefore the cleaner test, and the tracking-error measure guards against reading mechanical weight dispersion as economic disagreement.

APPENDIX C. ROBUSTNESS OF H2

C.1 Placebo tests (methodology)

The permutation placebo block-permutes the state-to-period assignment, preserving the regime autocorrelation and the per-state frequencies, and re-estimates the conditional-predictive-ability statistic; because a genuine state signal should not survive shuffling, the intended result is that the statistic collapses toward its no-information distribution. The wrong-mapping placebo inverts the economic map, routing each state to the method that performed worst rather than best within it, and the correct mapping is expected to outperform its inversion out of sample. The strong-benchmark test re-estimates the predictive-ability regression against the seven-control performance-only benchmark of Appendix B.2; rejection then means the state predicts the differential beyond recent relative performance at every horizon that benchmark spans. Results for the three placebos are reported in Exhibit C1.

C.2 Multiplicity and robustness design

One main specification is fixed on each design dimension, and the remainder are reported to demonstrate non-dependence rather than to optimize. The dimensions are: the universe (multi-asset main; equities-only and fixed-income-only for external validity); the regime detector (realized-volatility terciles main; the composite stress index and its MOVE-augmented variant); the trailing window of the volatility detector; the covariance estimator (Ledoit–Wolf shrinkage main; the sample covariance); and the transaction-cost level. None of the thresholds is optimized on out-of-sample results, and no turnover penalty enters the selection rule, so that implementation costs are an evaluation criterion rather than a tuned parameter. Exhibit C2 reports the results by regime detector and by volatility-detector window.

Multiple-testing correction is applied within the confirmatory family (H1 to H5) only; the detector, universe, volatility-window, and transaction-cost choices are robustness analyses and are not corrected across their Cartesian product.

C.3 Expected-return robustness and episode influence (methodology)

Two further checks isolate the source of the predictive content. First, because maximum-Sharpe is the only method that uses estimated expected returns, its state-dependence could reflect the instability of the sample mean rather than a broader property of estimation-dependent construction. We therefore re-estimate maximum-Sharpe with two shrinkage estimators of expected returns, both shrinking toward the cross-sectional grand mean with an intensity set by the Bayes–Stein rule (Jorion 1986) or the James–Stein rule (James and Stein 1961), holding the covariance estimator fixed at Ledoit–Wolf and the optimizer unchanged, so that only the expected-return input varies. Because full shrinkage toward a common mean turns maximum-Sharpe into minimum variance, Panel C of Exhibit 6 checks that the Bayes–Stein variant remains return-seeking. Second, we assess the influence of individual episodes on the pivotal test through a leave-one-episode-out exercise, re-estimating the conditional-predictive-ability statistics for the maximum-Sharpe pairs after removing each calendar year in turn (Exhibit C3) .

C.4 Results and exhibits

EXHIBIT C1. Placebo tests for H2

Panel A. Permutation placebo

	value
Actual, mean Wald	7.28
Permuted, mean Wald	2.24
p (mean Wald)	0.002
Actual, significant pairs	8
Permuted, significant pairs	1.4
p (n significant)	0.014

Panel B. Wrong-mapping placebo

	value
SR best-in-state (I_1)	0.90
SR worst-in-state (inverted)	0.12
Δ SR	0.78
Bootstrap p	< 0.001

Panel C. Strong-benchmark test (all method pairs)

Pair	Wald	p	sig 5%
1/N-MinVar	3.77	0.152	
1/N-MaxSharpe	13.39	0.001	*
1/N-HRP	4.87	0.088	
1/N-InvVol	3.46	0.177	
1/N-MaxDiv	7.16	0.028	*
1/N-HERC	2.72	0.256	
MinVar-MaxSharpe	10.07	0.006	*
MinVar-HRP	1.69	0.430	
MinVar-InvVol	2.49	0.288	

MinVar-MaxDiv	0.92	0.631	
MinVar-HERC	2.62	0.269	
MaxSharpe-HRP	8.93	0.011	*
MaxSharpe-InvVol	10.84	0.004	*
MaxSharpe-MaxDiv	9.78	0.008	*
MaxSharpe-HERC	9.62	0.008	*
HRP-InvVol	4.31	0.116	
HRP-MaxDiv	3.74	0.154	
HRP-HERC	1.80	0.406	
InvVol-MaxDiv	7.97	0.019	*
InvVol-HERC	1.63	0.443	
MaxDiv-HERC	2.40	0.301	

To guard against multiple testing across the family of pairwise comparisons (Exhibit 5), we apply the Holm–Bonferroni correction. All eight pairs that are significant at the 5% level individually remain significant after correction; in particular, all six pairs involving maximum-Sharpe survive, so the concentration of predictability in the most estimation-dependent method is not an artifact of multiplicity. The Benjamini–Hochberg procedure controlling the false discovery rate at 5% yields the same eight discoveries.

EXHIBIT C2. Robustness of H2 across detectors and volatility windows

Panel A. By detector.

Detector	Sig. pairs	MaxSharpe pairs	Mean Wald	Sharpe I ₁	Sharpe I ₀	Sharpe Div.	p (I ₁ vs I ₀)	p (Div vs I ₁)
<i>vol</i>	8	6	7.28	0.90	0.55	0.62	0.031	0.064
<i>composite</i>	2	0	2.88	0.56	0.55	0.62	0.954	0.692
<i>composite_m ove</i>	2	0	2.22	0.41	0.55	0.62	0.353	0.216

Panel B. By volatility-detector window

Vol window	Sig. pairs	MaxSharpe pairs	Mean Wald	Sharpe I ₁	Sharpe I ₀	Sharpe Div.	p (I ₁ vs I ₀)	p (Div vs I ₁)
9	4	2	4.74	0.79	0.55	0.62	0.133	0.181
13	8	6	7.28	0.90	0.55	0.62	0.031	0.064
17	9	6	10.03	0.95	0.55	0.62	0.047	0.053
26	6	6	5.38	0.84	0.55	0.62	0.146	0.235

NOTES: Panel A varies the state detector; Panel B the trailing window of the realized-volatility detector. "Sig. pairs": method pairs (of 21) with a significant state coefficient at 5%; "MaxSharpe pairs": those involving maximum-Sharpe (of 6). Sharpe ratios are annualized, out of sample, gross of costs; p-values use 4,999 bootstrap resamples (block length six; Appendix D).

EXHIBIT C3. Leave-one-episode-out: influence of individual years on the pivotal test

Year removed	Stress obs removed	Pairs significant (5%)	Mean Wald
— (baseline)	—	6 / 6	13.5
2011	4	6 / 6	13.4
2018	3	6 / 6	15.2
2020	7	6 / 6	10.9
2022	9	6 / 6	7.9
2023	8	6 / 6	13.7

Exhibit C4 reports the number of past observations available in each state at each formation date and the stability of the selected method. Because the selector learns on an expanding history, the training sample in each state grows through the period, and the minimum-eight-observation guard binds only in the first two years. Within each state the selected method appears stable rather than highly sensitive to sampling variation: outside the warm-up period the winner changes at fewer than 3% of formation dates in calm markets and never in stress. The modal selections are economically transparent, maximum-Sharpe in calm markets and equal weighting in stress, consistent with the mechanism hypothesized above.

EXHIBIT C4. Selector training sample and winner stability, by state.

Panel A. Past observations available in each state at each formation date (expanding history).

Statistic	Calm	Neutral	Stress
Mean	55.2	38.9	15.9
Median	63	37	10
10th percentile	8	13	1
25th percentile	29	17	9

Panel B. Stability of the selected method.

State	Decisions	P(method changes)	Excl. warm-up	Modal method (share)
Calm	90	6.7%	2.4%	Maximum-Sharpe (71%)
Neutral	82	9.9%	6.4%	1/N (60%)
Stress	43	7.1%	0.0%	1/N (81%)

NOTES: Panel A: the number of past formation dates classified in each state, at each of the 221 formation dates. Panel B: over the 215 selector decisions (221 less a six-rebalance burn-in), the frequency of a change in the selected method between consecutive same-state decisions, overall and excluding the warm-up, and the modal method.

Memory versus state. The state-conditional selector and performance-chasing differ in memory as well as in state-conditioning. To separate the two, we add a selector with the state-conditional selector's architecture (expanding history, minimum of eight observations, same fallback) that ignores the state and holds the method with the best mean performance over all past formation dates. Gross of costs, its Sharpe ratio is 0.64, against 0.90 for the state-conditional selector (difference 0.26, bootstrap $p = 0.10$) and 0.55 for performance-chasing (difference 0.09, $p = 0.59$). Net of costs the figures are 0.58, 0.83, and 0.46, with differences of 0.25 ($p = 0.13$) and 0.12 ($p = 0.47$). The state changes the selected method relative to this benchmark at 55% of formation dates. Longer memory alone thus lifts selection only to the level of diversification (0.59 net); descriptively, the state accounts for most of the remaining difference, although neither increment is individually significant.

APPENDIX D. SHARPE INFERENCE, MODEL CONFIDENCE SET, AND MULTIPLICITY

D.1 Inference methodology

Differences in Sharpe ratios are tested with the studentized circular block bootstrap, which accommodates serial dependence and the estimation of the Sharpe ratio (Ledoit and Wolf 2008). Where the honest object is a set of methods that cannot be statistically distinguished from the best, we report the model confidence set (Hansen, Lunde, and Nason 2011). The pivotal test of incremental information (H2) is the conditional-predictive-ability regression of Appendix B, and it is supported by three checks against spurious findings.

Throughout, the circular block bootstrap uses a block length of six formation dates and 4,999 resamples (2,000 for the reversal statistic in Exhibit 3, and 500 permutations for the permutation placebo). The robustness p-values in Exhibit C2 use the same Ledoit–Wolf test with independent bootstrap draws and may therefore differ slightly from those in Exhibits 7 and 9. Sharpe ratios in Exhibit 7 compound weekly returns at fixed target weights within each holding period, while Exhibit 9 uses the cost engine, in which weights drift within the period as drift-aware turnover requires; gross Sharpe ratios may therefore differ by 0.01 between the two exhibits. Heteroskedasticity-and-autocorrelation-robust statistics use a Newey–West lag of three in the conditional-predictive-ability regressions and four in the Sharpe-ratio difference tests and the model confidence set.

D.2 Estimation-stress diagnostics (exhibit)

EXHIBIT D1. Estimation-stress diagnostics by state

Panel A. Concentration ($1/N_{\text{eff}}$) and turnover, by method and state

Method	State	Concentration $1/N_{\text{eff}}$	Turnover
1/N	calm	0.083	0.009
1/N	neutral	0.083	0.010
1/N	stress	0.083	0.014
MinVar	calm	0.228	0.034
MinVar	neutral	0.264	0.037
MinVar	stress	0.264	0.051
MaxSharpe	calm	0.346	0.136
MaxSharpe	neutral	0.377	0.128
MaxSharpe	stress	0.479	0.150
HRP	calm	0.151	0.037
HRP	neutral	0.147	0.047
HRP	stress	0.150	0.042
InvVol	calm	0.098	0.010
InvVol	neutral	0.097	0.011
InvVol	stress	0.097	0.017
MaxDiv	calm	0.206	0.043
MaxDiv	neutral	0.182	0.048
MaxDiv	stress	0.196	0.064
HERC	calm	0.211	0.078

HERC	neutral	0.195	0.120
HERC	stress	0.224	0.099

Panel B. Covariance condition number, by state.

State	Condition number $\kappa(\Sigma)$
calm	68.0
neutral	67.5
stress	63.0

NOTES: Panel A: for each method and state, portfolio concentration ($1/N_{eff}$, the inverse of the effective number of assets) and one-period turnover. Panel B: the condition number of the estimated covariance matrix, by state.

REFERENCES

- Anderson, Evan W., and Ai-Ru (Meg) Cheng. 2022. "Portfolio Choices with Many Big Models." *Management Science* 68 (1): 690–715. <https://doi.org/10.1287/mnsc.2020.3876>.
- Ang, Andrew, and Geert Bekaert. 2002. "International Asset Allocation with Regime Shifts." *Review of Financial Studies* 15 (4): 1137–1187. <https://doi.org/10.1093/rfs/15.4.1137>.
- Ang, Andrew, and Allan Timmermann. 2012. "Regime Changes and Financial Markets." *Annual Review of Financial Economics* 4: 313–337. <https://doi.org/10.1146/annurev-financial-110311-101808>.
- Bae, Geum Il, Woo Chang Kim, and John M. Mulvey. 2014. "Dynamic Asset Allocation for Varied Financial Markets under Regime Switching Framework." *European Journal of Operational Research* 234 (2): 450–458. <https://doi.org/10.1016/j.ejor.2013.03.032>.
- Carhart, Mark M. 1997. "On Persistence in Mutual Fund Performance." *Journal of Finance* 52 (1): 57–82. <https://doi.org/10.1111/j.1540-6261.1997.tb03808.x>.
- Chopra, Vijay K., and William T. Ziemba. 1993. "The Effect of Errors in Means, Variances, and Covariances on Optimal Portfolio Choice." *Journal of Portfolio Management* 19 (2): 6–11. <https://doi.org/10.3905/jpm.1993.409440>.
- Choueifaty, Yves, and Yves Coignard. 2008. "Toward Maximum Diversification." *The Journal of Portfolio Management* 35 (1): 40–51. <https://doi.org/10.3905/JPM.2008.35.1.40>.
- Ciciretti, Vito, and Andrea Bucci. 2023. "Building Optimal Regime-Switching Portfolios." *The North American Journal of Economics and Finance* 64: 101837. <https://doi.org/10.1016/j.najef.2022.101837>.
- Clarke, Roger, Harindra de Silva, and Steven Thorley. 2013. "Risk Parity, Maximum Diversification, and Minimum Variance: An Analytic Perspective." *Journal of Portfolio Management* 39 (3): 39–53. <https://doi.org/10.3905/jpm.2013.39.3.039>.
- Cordoni, Francesco, and Alessio Sancetta. 2023. "Action-State Dependent Dynamic Model Selection." Working paper, Royal Holloway University of London. <https://doi.org/10.48550/arXiv.2307.04754>.
- Costa, Giorgio, and Roy H. Kwon. 2019. "Risk Parity Portfolio Optimization under a Markov Regime-Switching Framework." *Quantitative Finance* 19 (3): 453–471. <https://doi.org/10.1080/14697688.2018.1486036>.
- DeMiguel, Victor, Lorenzo Garlappi, and Raman Uppal. 2009. "Optimal versus Naive Diversification: How Inefficient Is the 1/N Portfolio Strategy?" *Review of Financial Studies* 22 (5): 1915–1953. <https://doi.org/10.1093/rfs/hhm075>.
- Garlappi, Lorenzo, Raman Uppal, and Tan Wang. 2007. "Portfolio Selection with Parameter and Model Uncertainty: A Multi-Prior Approach." *Review of Financial Studies* 20 (1): 41–81. <https://doi.org/10.1093/rfs/hhl003>.
- Giacomini, Raffaella, and Halbert White. 2006. "Tests of Conditional Predictive Ability." *Econometrica* 74 (6): 1545–1578. <https://doi.org/10.1111/j.1468-0262.2006.00718.x>.
- Grunnesjö, Marcus, David Islip, Xuanze Li, Yanni Lu, and Roy H. Kwon. 2026. "A Regime-Switching Decision Support System for Dynamic Portfolio Optimization." *Annals of Operations Research*. <https://doi.org/10.1007/s10479-025-06990-1>.

- Hansen, Peter R. 2005. "A Test for Superior Predictive Ability." *Journal of Business & Economic Statistics* 23 (4): 365–380. <https://doi.org/10.1198/073500105000000063>.
- Hansen, Peter R., Asger Lunde, and James M. Nason. 2011. "The Model Confidence Set." *Econometrica* 79 (2): 453–497. <https://doi.org/10.3982/ECTA5771>.
- Hautsch, Nikolaus, and Stefan Voigt. 2019. "Large-Scale Portfolio Allocation under Transaction Costs and Model Uncertainty." *Journal of Econometrics* 212 (1): 221–240. <https://doi.org/10.1016/j.jeconom.2019.04.028>.
- Jagannathan, Ravi, and Tongshu Ma. 2003. "Risk Reduction in Large Portfolios: Why Imposing the Wrong Constraints Helps." *Journal of Finance* 58 (4): 1651–1684. <https://doi.org/10.1111/1540-6261.00580>.
- James, W., and Charles Stein. 1961. "Estimation with Quadratic Loss." In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, 361–379. Berkeley: University of California Press.
- Jorion, Philippe. 1986. "Bayes–Stein Estimation for Portfolio Analysis." *Journal of Financial and Quantitative Analysis* 21 (3): 279–292. <https://doi.org/10.2307/2331042>.
- Kan, Raymond, and Guofu Zhou. 2007. "Optimal Portfolio Choice with Parameter Uncertainty." *Journal of Financial and Quantitative Analysis* 42 (3): 621–656. <https://doi.org/10.1017/S0022109000004129>.
- Kritzman, Mark, Sébastien Page, and David Turkington. 2012. "Regime Shifts: Implications for Dynamic Strategies." *Financial Analysts Journal* 68 (3): 22–39. <https://doi.org/10.2469/faj.v68.n3.3>.
- Ledoit, Olivier, and Michael Wolf. 2004. "A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices." *Journal of Multivariate Analysis* 88 (2): 365–411. [https://doi.org/10.1016/S0047-259X\(03\)00096-4](https://doi.org/10.1016/S0047-259X(03)00096-4).
- Ledoit, Olivier, and Michael Wolf. 2008. "Robust Performance Hypothesis Testing with the Sharpe Ratio." *Journal of Empirical Finance* 15 (5): 850–859. <https://doi.org/10.1016/j.jempfin.2008.03.002>.
- López de Prado, Marcos. 2016. "Building Diversified Portfolios That Outperform Out-of-Sample." *Journal of Portfolio Management* 42 (4): 59–69. <https://doi.org/10.3905/jpm.2016.42.4.059>.
- Markowitz, Harry. 1952. "Portfolio Selection." *Journal of Finance* 7 (1): 77–91. <https://doi.org/10.2307/2975974>.
- Pástor, Luboš. 2000. "Portfolio Selection and Asset Pricing Models." *Journal of Finance* 55 (1): 179–223. <https://doi.org/10.1111/0022-1082.00204>.
- Raffinot, Thomas. 2018a. "Hierarchical Clustering-Based Asset Allocation." *Journal of Portfolio Management* 44 (2): 89–99. <https://doi.org/10.3905/jpm.2018.44.2.089>.
- Raffinot, Thomas. 2018b. "The Hierarchical Equal Risk Contribution Portfolio." *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3237540>.
- Timmermann, Allan. 2006. "Forecast Combinations." In *Handbook of Economic Forecasting*, vol. 1, edited by Graham Elliott, Clive W. J. Granger, and Allan Timmermann, 135–196. Amsterdam: Elsevier.

Tu, Jun, and Guofu Zhou. 2011. “Markowitz Meets Talmud: A Combination of Sophisticated and Naive Diversification Strategies.” *Journal of Financial Economics* 99 (1): 204–215. <https://doi.org/10.1016/j.jfineco.2010.08.013>.

White, Halbert. 2000. “A Reality Check for Data Snooping.” *Econometrica* 68 (5): 1097–1126. <https://doi.org/10.1111/1468-0262.00152>.