

Decay-Aware Cluster-Coherent Alpha & Risk

One characteristic map for time-varying alpha and portfolio risk

Siqi (Kristy) Jia (sj927) · Patrick Liu (tl892) · Wendy Liu (jl4585) · Oscar Chau (cc3532)

MGT 924 Statistical Foundations · Professor Theis Ingerslev Jensen · Yale School of Management · October 2026

ABSTRACT The 402 firm characteristics in the competition data are highly redundant, and the predictive value of their common components need not be stable through time. We cluster the characteristics into 25 data-driven themes (Figure 1) and use the same theme map on both sides of portfolio construction. An ensemble of linear and nonlinear learners first produces stock-level return forecasts. We decompose those forecasts into common theme and stock-specific components, then regularise the common component with shrinkage estimates of theme premia formed under several memory lengths and blended using prior within-refit walk-forward performance under the current theme map. The same theme exposures define the factor structure of the portfolio covariance matrix. The resulting mean-variance portfolio therefore adapts to changing theme premia while explicitly controlling the risk of those same bets. Every transformation, tuning decision and portfolio weight uses only information available at the formation date. Over the competition's test period (Dec 1989–Nov 2023) the strategy earns a Sharpe ratio of 2.51 (2.41 without the decay component), against 0.54 for the market and 1.44 for the problem-set baseline, with a market beta of 0.17.



1 Methodology

1.1 Timing and information set

Let t index month-ends, i stocks and j characteristics. The portfolio formed at the end of month t earns the excess return $r_{i,t+1}$ and may only use

$$\mathcal{I}_t = \{x_{ij,s} : s \leq t\} \cup \{r_{i,s+1} : s \leq t-1\} \cup \{\text{daily returns up to day } t\}. \quad (1)$$

The label of month $t-1$ (the return over month t) is the most recent one the models learn from. Every step below uses $\mathcal{I}_t$ alone: themes, models, penalties and memory weights are re-estimated each December, theme premia every month.

1.2 Step 1 – Make characteristics comparable

Each month, every characteristic is replaced by its centred cross-sectional percentile, with missing values at the median, $\tilde{x}_{ij,t} = (\text{rank}_t(x_{ij,t}) - 1) / (N_{j,t} - 1) - \frac{1}{2} \in [-\frac{1}{2}, \frac{1}{2}]$, and $\tilde{x}_{ij,t} = 0$ if missing. The target $\tilde{y}_{i,t}$ is the normal score of the cross-sectional rank of $r_{i,t+1}$. A long-short portfolio needs only the *ordering* of stocks within a month, and ranking stops a handful of extreme returns from dominating the fit.

1.3 Step 2 – Let the data find the themes

Over the trailing ten years we estimate rank correlations ρ_{jk} between characteristics on a seeded sample of at most 200,000 stock-months and cluster them by average linkage on the distance $d_{jk} = 1 - |\rho_{jk}|$, cutting the tree at 0.5 (at most 25 themes), as in Jensen, Kelly & Pedersen (2023). Signs are aligned with each theme's first principal component, $s_j = \pm 1$, and a stock's exposure to theme c is the z-scored signed average of its member ranks,

$$e_{ic,t} = z_t \left(\frac{1}{|c|} \sum_{j \in c} s_j \tilde{x}_{ij,t} \right). \quad (2)$$

The latest fit finds 25 themes, from 1 to 210 characteristics each; we name each after its most central (medoid) member.

THE IDEA: ONE MAP FOR BOTH PREDICTION AND RISK

Statistical risk factors, such as principal components of returns, need not line up with the characteristic bets that a forecast makes. Here the themes enter twice. On the alpha side, the part of the forecast that the themes span is re-estimated from the themes' own return history, with a memory chosen by its track record. On the risk side, the same themes are the factors, so the optimiser recognises and risk-adjusts the common theme exposures embedded in the forecast, and keeps a theme bet only to the extent that its forecast pays for its estimated risk. Section 3.1 tests this.

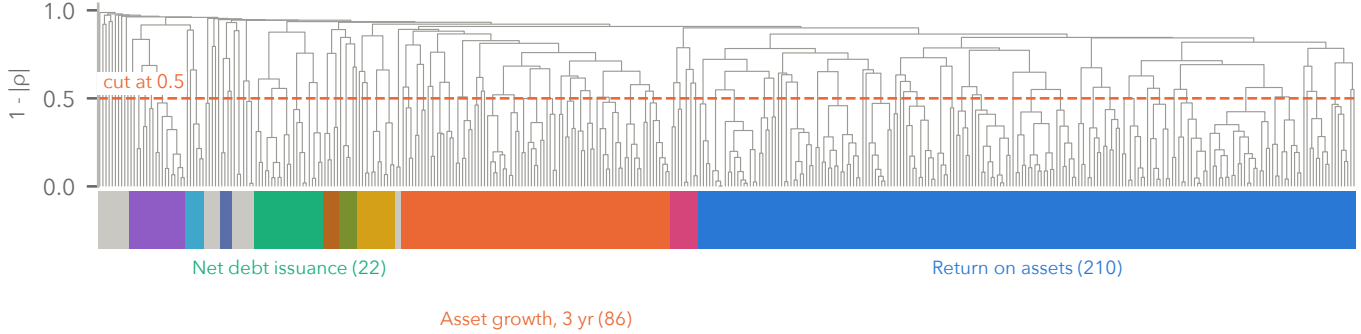


Figure 1. The map. Average-linkage tree of all 402 characteristics on $1 - |\rho|$ (latest refit, trailing ten years). Each coloured block is one theme; the ten largest are coloured and those with 20 or more members are named after their medoid (members in parentheses). 402 published signals collapse into 25 distinct bets, and the largest theme alone absorbs 210 of them.

1.4 Step 3 – Forecast returns with three complementary learners

Ridge on all ranked characteristics, $\hat{\beta}_t(\alpha) = (X_t^\top X_t + \alpha \bar{\lambda} I)^{-1} X_t^\top \tilde{y}_t$, where $\bar{\lambda}$ is the average eigenvalue of $X_t^\top X_t$. **Random-feature ridge** applies the same estimator to $\phi(\tilde{x}) = P^{-1/2}[\sin(\Omega^\top \tilde{x}), \cos(\Omega^\top \tilde{x})]$ with $P = 1,000$ Gaussian directions Ω , a cheap, high-dimensional way to capture smooth non-linearities (Kelly, Malamud & Zhou, 2024). For both ridges the penalty is chosen inside the code at every refit: the model is fit on all but the last five years of the training sample, and the penalty kept is the one with the highest mean monthly information coefficient (IC) over those five years, $\alpha_t^* = \arg \max_{\alpha \in \{10^{-4}, \dots, 10^4\}} \overline{\text{IC}}_{[t-60, t)}(\alpha)$; the model is then refit on the full sample. **Gradient-boosted trees** capture thresholds and interactions, with 300 shallow trees (31 leaves, at least 2,000 stocks per leaf), a learning rate of 0.05, an ℓ_2 penalty and random halves of the features. The random-feature ridge and the trees train on seeded subsamples (400k and 600k stock-months). The forecast is the equal-weight average, $\mu_{i,t} = z_t(\frac{1}{3}[z_t(\hat{y}_{i,t}^{\text{ridge}}) + z_t(\hat{y}_{i,t}^{\text{RFF}}) + z_t(\hat{y}_{i,t}^{\text{trees}})])$; fixed equal weights avoid estimating combination weights, another source of error.

1.5 Step 4 – Make the common theme component decay-aware

Let E_t be the $N_t \times K_t$ matrix of theme exposures. The ML forecast splits into the part the themes span and a stock-specific residual, $\mu_t^{\text{ML}} = \mu_t^{\text{theme}} + \mu_t^{\text{res}}$, with $\mu_t^{\text{theme}} = E_t(E_t^\top E_t + \epsilon I)^{-1} E_t^\top \mu_t^{\text{ML}}$ and a negligible numerical ridge ϵ . Only the theme part is modified. Each theme defines a portfolio formed on the whole formation-date universe, $v_{i,c,s} = 2e_{i,c,s} / \sum_i |e_{i,c,s}|$, whose return is $f_{c,s+1} = \sum_i v_{i,c,s} r_{i,s+1}$; a missing return counts as zero after the weights are formed, with no renormalisation. For half-lives $h \in \{1, 3, 10, 30 \text{ years}, \infty\}$ the premia are exponentially weighted means of f shrunk toward zero,

$$b_{c,t}^{(h)} = \frac{n_{\text{eff}}^{(h)}}{n_{\text{eff}}^{(h)} + 60} \cdot \frac{\sum_{s < t} \delta_h^{t-1-s} f_{c,s+1}}{\sum_{s < t} \delta_h^{t-1-s}}, \quad \delta_h = 2^{-1/(12h)}, \quad (3)$$

where n_{eff} is the effective number of months, so short memories are shrunk hardest. A within-refit walk-forward ledger records each memory's subsequent return, $\rho_{s+1}^{(h)} = b_s^{(h)\top} f_{s+1} / \|b_s^{(h)}\|_1$, with each premium estimate using only earlier theme returns under the current theme map. Every December the memories get weights $\pi_h \propto \max(\text{SR}_{[t-120, t)}^{(h)}, 0)$ from the last ten years of that ledger, and

$$\mu_t^{\text{decay}} = E_t \sum_h \pi_h b_t^{(h)}, \quad \mu_t^{\text{final}} = \mu_t^{\text{res}} + (1 - \gamma) \mu_t^{\text{theme}} + \gamma \mu_t^{\text{decay}}, \quad (4)$$

with μ_t^{decay} rescaled to the cross-sectional dispersion of μ_t^{theme} , so that γ trades off two forecasts of the same size. At each refit the theme history is rebuilt under the current themes, using at every past month only returns realised by then. The stock-specific ML forecast is never touched, and $\gamma = 0$ recovers the plain ML forecast.

1.6 Step 5 – Risk model and portfolio weights

Before risk estimation, daily returns are clipped to $[-50\%, 100\%]$, and observations with missing daily returns are omitted. For each day τ in the 24 months before t , a cross-sectional regression of daily excess returns on $B_t = [\mathbf{1}, E_t]$, a

common factor plus the same theme exposures at the preceding month-end, gives factor returns f_τ and residuals $\varepsilon_{i\tau}$:

$$r_{i\tau} = b_i^\top f_\tau + \varepsilon_{i\tau}, \quad \Sigma_t = B_t F_t B_t^\top + D_t, \quad w_t = c_t \Sigma_t^{-1} \mu_t^{\text{final}}. \quad (5)$$

F is the exponentially weighted covariance of f_τ (half-life 252 days). D holds exponentially weighted residual variances (half-life 63 days), shrunk toward the cross-sectional median with the weight of 20 observations. The Woodbury identity reduces Σ_t^{-1} to a $(K+1) \times (K+1)$ inverse. Positions are capped at ten times the average absolute weight, and c_t scales predicted volatility to 10% a year. The portfolio holds all stocks available each month.

ASSUMPTIONS WE MAKE EXPLICIT

A1. `main(chars, features, daily_ret)` receives the three tables of the data guide and returns `id` (integer), `eom`, `w` for every `ctff_test` month. **A2.** The target is `ret_exc_lead1m`, market equity is `market_equity`, and the model uses exactly the 402 characteristics listed in `ctff_features`. **A3.** A missing next-month return counts as zero, the leaderboard rescales returns to 10% volatility, and benchmark Sharpe ratios are taken as published on the leaderboard.

2 Validation

2.1 Identification strategy

TRAINING, VALIDATION, TEST, AND WHAT WAS FROZEN

Training history: months before 1975 are used only to fit models. **Validation:** Jan 1975–Nov 1989 (179 months), on which the decay weight $\gamma \in \{0, 0.25, 0.5, 0.75, 1\}$ was chosen, as the value with the highest validation Sharpe ratio of the full model, and each component is assessed; the other settings in Section 1 were fixed in advance, not tuned. **Test:** the months flagged `ctff_test`, Dec 1989–Nov 2023, evaluated once with γ frozen. **Frozen:** every setting in Section 1 and $\gamma = 1.00$. **Chosen inside the code at every refit, from past data only:** the themes, the ridge penalties and the memory weights π_h . **Disclosure:** an earlier version of this model, identical except for the decay component ($\gamma = 0$), had been evaluated on the test period before the decay component was designed. The component's settings (half-lives, prior of 60 months, ten-year ledger) were fixed before it was run, γ was chosen on validation data only, and no test-period result was used to choose anything else.

Table 1 isolates each component on the validation period; every row is scaled to 10% volatility with the same risk model, and intervals come from a stationary bootstrap (mean block 6 months, 2,000 draws).

Table 1. Validation period, Jan 1975–Nov 1989: what each component adds. Rows 3–4 and 6 use the frozen $\gamma = 1.00$. Turnover is half the sum of absolute weight changes per unit of gross exposure per month.

Validation design, 1975–1989	Sharpe	90% interval	Beta	Turnover
Problem-set ridge baseline (decile long-short)	2.84	[2.43, 3.31]	−0.36	62%
ML ensemble, rank weights	3.24	[2.86, 3.70]	−0.35	48%
ML residual + fixed full-history theme premia	4.79	[4.31, 5.37]	0.03	50%
ML residual + decay-aware theme premia	4.16	[3.71, 4.70]	−0.03	48%
ML + theme risk model, no decay	5.68	[5.07, 6.45]	0.08	53%
Decay-aware CCAR, full model	6.20	[5.53, 7.04]	0.02	53%

Table 1 answers three questions separately. *Does machine learning help?* The ensemble with rank weights earns 3.24 against 2.84 for the baseline ($p < 0.01$). *Does decay-aware theme prediction help?* With rank weights, decay-aware theme premia raises the Sharpe ratio from 3.24 to 4.16 ($p < 0.01$), but fixed full-history premia do better still (4.79), so with rank weights the gain comes from re-estimating the theme premia from the themes' own returns, not from forgetting. *Does cluster-coherent risk management help?* The theme risk model raises the ML forecast's Sharpe ratio from 3.24 to 5.68 ($p < 0.01$) and moves market beta from −0.35 to 0.08. Combined, the selected full model earns **6.20** against 5.68 without the theme-premia overlay. Because γ was selected on this validation sample, this difference is descriptive rather than an independent significance test. Bootstrap p -values are one-sided, from the same resampled months for both designs.

Choosing γ . Validation Sharpe ratio of the full model for each γ (0.00: 5.68, 0.25: 5.84, 0.50: 5.98, 0.75: 6.10, 1.00: 6.20). The maximum, $\gamma = 1.00$, was frozen in the code before any test-period month of this version was evaluated. It is the edge of the grid: the ML forecast's theme part is replaced entirely, its stock-specific part kept. We did not extend the grid, since a weight above one would bet against the learners' own theme forecast.

2.2 Integrity of the evidence

- **Timing.** Ranks use only the current cross-section; a return enters training, the theme ledger or the memory weights only once realised; every refit and tuning choice uses earlier months only; daily returns end on the portfolio date. Stocks are never dropped for a missing next-month return, which would reveal delistings. **No hand-picking.** All 402

- characteristics enter every model; themes, penalties, memory weights and portfolio weights are chosen by algorithms inside the code.
- **Look-ahead tests.** Re-running `main()` on data cut off at Jun 1995 and Dec 2005, with the cut-off month's next-month returns also deleted, reproduces all 135,174 and 439,387 earlier weights exactly (maximum difference 0). A full run of `main()` and the independent backtest run give identical weights in all 885,698 test-period positions. Seeds derive from the refit year and package versions are pinned.

3 Results

Over the 408 test months the Sharpe ratio is 2.51 (standard error 0.19; 90% bootstrap interval 1.96–3.13) against 0.54 for the market, with a market beta of 0.17 and a CAPM alpha t -statistic of 9.0. At 10% volatility the largest fall in compounded wealth is –20.9%, against –37.4% for the market. The problem-set baseline earns 1.44 over the same months ($p < 0.01$ for the difference). The Sharpe ratio is 5.84 in 1990–99 and 1.03 in 2020–23, so performance has weakened over time. Stocks outside the large and mega-cap groups provide 63% of the return; the large and mega-cap positions alone earn a Sharpe ratio of 1.44.

Table 2. Test period, Dec 1989–Nov 2023 (408 months), at 10% annual volatility as on the leaderboard. Standard errors: Lo (2002). Market: value-weighted, from the competition data.

Statistic	Strategy	Market
Sharpe ratio (standard error)	2.51 (0.19)	0.54 (0.17)
90% bootstrap interval, Sharpe ratio	[1.96, 3.13]	[0.23, 0.88]
Average annual return at 10% volatility	25.1%	5.4%
Maximum drawdown of wealth at 10% volatility	–20.9%	–37.4%
Worst month at 10% volatility	–17.4%	–11.9%
Share of positive months	81%	63%
Skewness of monthly returns	–0.47	–0.57
Market beta (t -statistic)	0.17 (0.9)	1.00
CAPM alpha t -statistic	9.0	–
Sharpe ratio, 1990–99	5.84	0.96
Sharpe ratio, 2000–09	2.36	–0.06
Sharpe ratio, 2010–19	1.79	0.99
Sharpe ratio, 2020–23	1.03	0.57

Table 3. Sharpe ratio, 1990–2023, against the competition's benchmark portfolios (published on the leaderboard).

Portfolio	Sharpe
Markowitz-ML	2.52
Decay-aware CCAR	2.51
Instrumented PCA factor portfolio	1.95
Problem-set baseline (ridge, deciles)	1.44
Kozak-Nagel-Santosh rank-weighted factors	1.42
Minimum variance	0.95
Factor-ML	0.89
Equal-weight portfolio	0.49

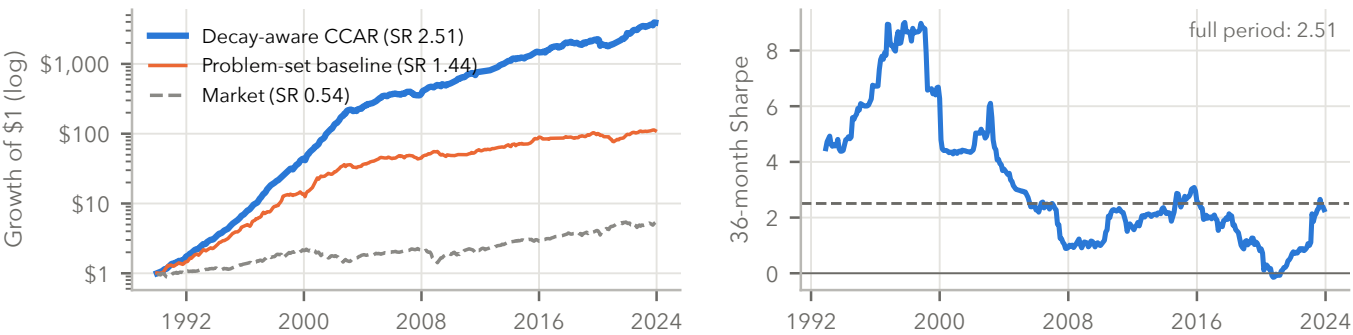


Figure 2. Test period. Left: growth of \$1 with monthly returns compounded, every strategy scaled to 10% annual volatility (log scale). Right: Sharpe ratio over rolling 36-month windows; the dashed line is the full-period value.

3.1 Does one map really help? Where the return and the risk come from

On average 41% of the cross-sectional variance of the ML forecast lies in the span of the themes, so a large part of what the learners predict is a theme bet. Splitting each portfolio's return into common, theme and stock-specific parts (Figure 3): with rank weights, theme bets earn 5.2% a year at 10% volatility and contribute 56% of the variance. The full model cuts theme exposure per unit of gross position from 3.34 to 0.97 and the theme share of variance to 3%, while the stock-specific return moves from 12.5% to 22.6%; that component alone has a Sharpe ratio of 2.31. The optimiser does not remove theme bets, it prices them: exposure survives where the forecast pays for its risk.

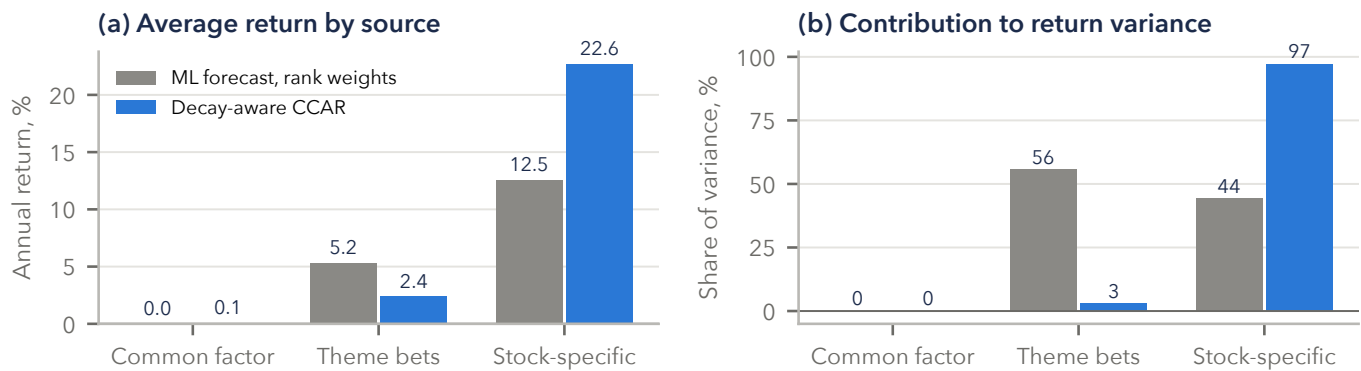


Figure 3. Return and risk attribution, test period. Each month realised returns are regressed on the common factor and the theme exposures used to build the portfolio. (a) Average annual return of each part, scaled to 10% volatility. (b) Each part's covariance with the total return as a share of its variance (shares add to 100%). Grey: ML forecast with rank weights; blue: the full model.

3.2 Does decay awareness help?

Figure 4(a) shows how the ledger weights the five memories. Over the test period the average weights are full history 27%, 30 yr 25%, 10 yr 22%, 3 yr 15%, 1 yr 11%. Panel (b) cumulates the full model's return minus the no-decay model's, both at 10% volatility. Over the test period the difference averages 1.0% a year; the two portfolios' returns correlate at 0.997. The test-period Sharpe ratio is 2.51 with the overlay and 2.41 without it (bootstrap $p < 0.01$), and 1.44 against 1.40 since 2010. This difference measures the theme-premia overlay as a whole. Two facts say the gain owes more to re-estimating theme premia from theme returns than to forgetting: with rank weights, fixed full-history premia beat decay-aware ones on validation (Table 1), and the ledger itself leans toward long memories, putting 52% on 30 years or full history. Decay awareness mainly lets the data choose how much history to trust, and here it chose to remember.

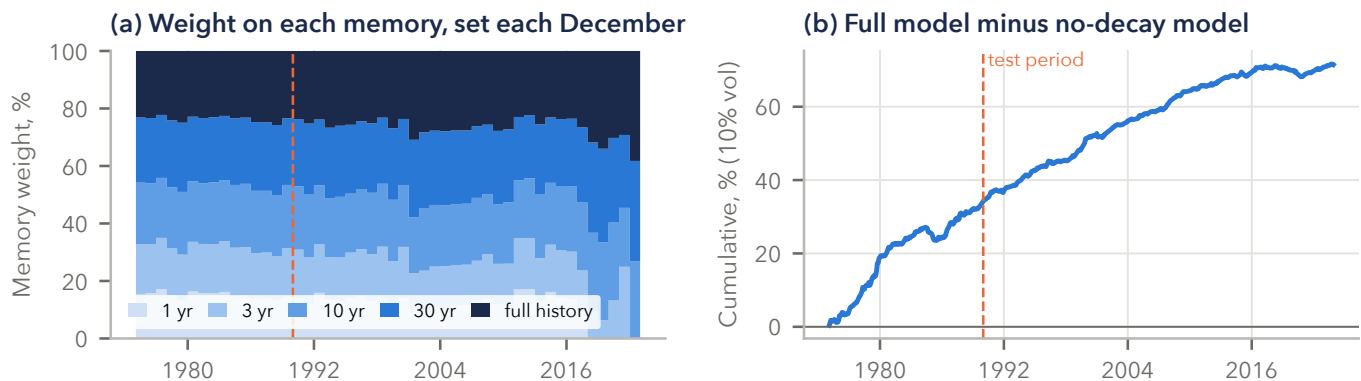


Figure 4. Decay awareness, 1975–2023. (a) Memory weights π_{it} chosen each December from the preceding ten years of the within-refit walk-forward ledger. (b) Cumulative difference between the monthly returns of the full model and of the same model with $\gamma = 0$, both scaled to 10% volatility. The dashed line marks the start of the test period.

4 What an investor should expect

- **Expected performance.** The full-sample Sharpe ratio of 2.51 should not be read as a forward expectation. Performance has weakened over time, and the recent subsample is both lower and estimated imprecisely. We therefore regard the post-2010 Sharpe ratio as more informative about prospective performance: 1.44 over 167 months, with a 90% bootstrap interval of 0.86–2.19. Trading and financing costs would further reduce realised performance.
- **Costs and leverage.** On the leaderboard's 10%-realized-volatility scale, the gross book is $4.0\times$ capital and monthly trading is about 47% of the gross position. The edge disappears at about 56 bp of cost per dollar traded, and would shrink most in small stocks.
- **Risk forecasts.** Weights scaled to 10% predicted volatility realised 32%: like most optimised portfolios, the risk model is optimistic, so positions should be sized on realised volatility.

Disclosure. No external code is borrowed. The code was written with the help of an AI assistant (Claude); we reviewed it and are responsible for it.

References. Jensen, Kelly & Pedersen (2023), Is there a replication crisis in finance?, *J. Finance* 78(5). Jensen, Kelly, Malamud & Pedersen (2024), Machine learning and the implementable efficient frontier, working paper. Kelly, Malamud & Zhou (2024), The virtue of complexity in return prediction, *J. Finance* 79(1). Lo (2002), The statistics of Sharpe ratios, *Financial Analysts J.* 58(4). McLean & Pontiff (2016), Does academic research destroy stock return predictability?, *J. Finance* 71(1). Politis & Romano (1994), The stationary bootstrap, *JASA* 89(428).