

MGT 924

Statistical Foundations

Problem Set 3

Three Return Signals, One Risk Model:
A JKP Common Task Framework Submission

Antoine Boisard

Anika Bokka

Connor Hipp

Sunmeet Gulati

Yale School of Management

Professor Theis Ingerslev Jensen

Fall 2026

Summary

We submit a long-short portfolio of about 2,068 liquid US stocks, rebalanced monthly and close to dollar-neutral (median net exposure 1% of capital). Three models predict each stock's next-month return from 402 characteristics; a Barra-style factor risk model turns their average into mean-variance weights; and the size of the whole position is scaled each month by the volatility the strategy recently realised.

Track record. Over the scored test period, December 1989 to November 2023 (408 months), the annualised Sharpe ratio is **2.79** (95% interval 2.40 to 3.17), 79% of months are positive, realised volatility is 13% against a 10% target, and the worst drawdown at 10% volatility is -8.2%. That would place 16th of 57 on the public leaderboard as of 7 October 2026, above all six benchmark models (the best, Markowitz-ML, scores 2.52).

What an investor should expect. Less than the headline. The Sharpe ratio is 4.81 in 1990–2004, 1.62 since 2005 and 1.14 since 2015, and every figure is before trading costs. At 10 basis points per dollar traded, the test Sharpe ratio falls to 2.20. After realistic costs, an investor should plan for a Sharpe ratio below 1.

Data, Target and Timing

The competition data hold 1,369,324 stock-months from December 1951 to November 2023 with 402 characteristics; the 885,698 stock-months from December 1989 onward are scored. Each month we replace characteristic k of stock i by its cross-sectional rank among the $n_{k,t}$ stocks where it is observed,

$$x_{i,k,t} = \frac{\text{rank}(c_{i,k,t}) - 1}{n_{k,t} - 1} - \frac{1}{2} \in \left[-\frac{1}{2}, \frac{1}{2}\right], \quad (1)$$

and set a missing value to 0, the rank of the median stock. The target $y_{i,t+1}$ is the next-month excess return, winsorised at the 1st and 99th cross-sectional percentiles and demeaned within the month, so the models learn which stocks beat others rather than the direction of the market.

Timing rule. A portfolio formed at the end of month t uses only information available at t : every model is refit each December on an expanding window that admits a stock-month only once its return month has ended.

Return Signals

Linear ridge (Sunmeet). On the ranked characteristics X , $\hat{\beta}_\lambda = (X'X + \lambda I)^{-1}X'y$, with λ chosen at each refit by out-of-sample error over the last ten years of the training window.

Random Fourier features (Sunmeet). To capture nonlinearities cheaply, the ranks are mapped into 512 features

$$\phi(x) = [\sin(W'x), \cos(W'x)], \quad W_{jk} \sim \mathcal{N}\left(0, \frac{1}{P}\right), \quad (2)$$

with $P = 402$ and W drawn once with a fixed seed, and the same ridge is fit on $\phi(x)$.

Gradient-boosted trees (Connor). Trees can learn that a characteristic matters only among small stocks or only at the extremes, which ridge cannot. LightGBM builds the prediction as a sum

of M shallow trees,

$$\hat{f}(x) = \sum_{m=1}^M \eta h_m(x), \quad h_m = \arg \min_h \sum_i (y_i - \hat{f}_{m-1}(x_i) - h(x_i))^2 + \Omega(h), \quad (3)$$

where each tree fits the residuals of the previous ones and Ω is an L2 penalty on the leaf values. It trains on a fixed half of the stock-months, chosen by a hash of stock and month so that truncating the data never changes past samples. The settings were fixed before any results were computed (31 leaves, $\eta = 0.05$, half the features and rows per tree, at least 500 stocks per leaf); the data choose only M , by early stopping on the last five years of each training window. M is 53–303 between 1990 and 2003 and at most 23 after 2005, consistent with less structure being left to find.

Theme portfolios (Antoine). Each characteristic becomes a dollar-neutral long-short factor with next-month return $F_{k,t+1}$. Every December, using past data only, factors are oriented so their past mean is positive and grouped into 13 themes by Ward clustering of their correlations. Factors whose past mean is not significant,

$$|t_k| = \frac{|\bar{F}_k|}{\hat{\sigma}_k/\sqrt{T}} < 1, \quad (4)$$

are left out, tangency weights are estimated on the 13 theme returns, and the result is mapped back to stocks. This signal was a candidate for the blend; the next sections show how it was judged.

Ensemble. Each signal is z-scored within the month and the selected signals are averaged with equal weights,

$$\mu_{i,t} = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \frac{p_{i,t}^s - \bar{p}_t^s}{\text{sd}_t(p^s)}, \quad \mathcal{S} = \{\text{ridge, RFF, LightGBM}\}. \quad (5)$$

Equal weights require no estimation, which makes them hard to overfit.

From Predictions to a Portfolio

Risk model and weights (Sunmeet). Following Barra, $r_t = E_t f_t + \varepsilon_t$, so

$$\Sigma_t = E_t F_t E_t' + \text{diag}(v_t), \quad (6)$$

where E_t holds 12 industry dummies (Fama-French 12) and 13 style factors from Ward clustering of characteristic correlations, F_t is the exponentially weighted covariance of daily factor returns (half-life 252 days), and v_t the exponentially weighted residual variances (half-life 63 days). The base weights are the tangency portfolio at 10% ex-ante volatility,

$$w_t^0 = \kappa_t \Sigma_t^{-1} \mu_t, \quad \kappa_t \text{ such that } \sqrt{12 w_t^{0'} \Sigma_t w_t^0} = 10\%, \quad (7)$$

computed with the Woodbury identity so we never invert a $2,000 \times 2,000$ matrix.

Volatility management (Antoine). The risk model predicts 10% but the base portfolio realised 26%: a known reason is that an optimiser puts the most weight where estimated risk is lowest, so its own risk forecast is biased down (Michaud, 1989). Following Moreira and Muir (2017), we scale by realised rather than predicted risk. The daily return of last month's base weights and its realised variance are

$$R_d = \sum_i w_{i,t-1}^0 r_{i,d}, \quad RV_t = \sum_{d \in t} R_d^2, \quad \hat{\sigma}_t^2 = \frac{1}{6} \sum_{m=0}^5 RV_{t-m}, \quad (8)$$

and the submitted weights are

$$w_{i,t} = w_{i,t}^0 \cdot \frac{0.10/\sqrt{12}}{\hat{\sigma}_t}. \quad (9)$$

The multiplier is common to all stocks, so positions keep their direction; the strategy holds less after turbulent months and more after calm ones, using only daily returns dated on or before t .

How We Guarded Against Overfitting

- **Validation and test kept apart.** Every choice made in combining the models used only the validation period, December 1961 to November 1989 (336 months in which every signal exists). Antoine chose the $|t| \geq 1$ cutoff and the six-month window on 1962–1989; we did not re-tune them.
- **Rules written down before each run.** R1: start from all four signals and drop any whose removal raises the validation Sharpe ratio. R2: turn volatility management on if it raises the validation Sharpe ratio of the R1 blend. Under R1, removing the theme signal raised validation from 5.22 to 5.24, while removing ridge, RFF or LightGBM lowered it (5.21, 5.14, 5.01), so the blend is ridge + RFF + LightGBM. Under R2, volatility management raised it from 5.24 to 5.43, so it is on.
- **The competition’s look-ahead test, run ourselves.** We re-ran the full pipeline on data truncated at December 1995 and at December 2008. Every weight up to the truncation date is identical to the full run: the maximum difference is 0 across 150,073 test stock-months for 1995 and 0 across 509,008 for 2008. The code is deterministic (fixed seeds, fixed thread count, LightGBM’s deterministic mode), and our R version of Antoine’s model reproduces his Python output to within 10^{-15} .

Results

Table 1 reports the annualised Sharpe ratio $\widehat{SR} = \sqrt{12} \bar{r} / \hat{\sigma}_r$ for each set of signals, with and without volatility management (VM).

Three results stand out. First, the blend beats each of its parts in both periods; its monthly rank correlation with next-month returns averages 0.052 in the test period, against 0.049 for ridge, 0.048 for RFF and 0.045 for LightGBM. Second, volatility management raises the validation Sharpe ratio of every risk-model row. On the test period it leaves our Sharpe ratio essentially unchanged (2.79 to 2.79) but brings realised volatility from 26% down to 13%, much closer to the 10% target, and the worst drawdown at 10% volatility from -9.9% to -8.2%: the portfolio now carries the risk it advertises. Third, on the test period all four signals with VM score 2.85, above our 2.79. We kept the blend our rule chose, because switching on a test-period difference is exactly the overfitting we set out to avoid.

Why It Works

Our forecasts are modest: the blend’s monthly rank correlation with next-month returns is about 0.05. The edge is in what we do with them. The same blended signal earns 1.12 as a plain rank

Table 1. Annualised Sharpe ratio by signal set. All rows except the last two use the risk model (6)–(7); “VM” adds (8)–(9).

Signals in the average	Validation, 1961–1989		Test, 1989–2023	
	no VM	VM	no VM	VM
Ridge + RFF + LightGBM (submitted: VM)	5.24	5.43	2.79	2.79
All four, including themes	5.22	5.40	2.81	2.85
Ridge + RFF (Sunmeet’s original model)	5.06	5.23	2.61	2.57
LightGBM only	5.14	5.31	2.61	2.63
Ridge only	4.97	5.14	2.53	2.51
RFF only	4.95	5.06	2.53	2.49
Themes only	3.40	3.47	1.87	2.05
Antoine’s stand-alone theme portfolio	2.90	3.08	1.81	2.01
Submitted signal, rank long-short	2.66	–	1.12	–

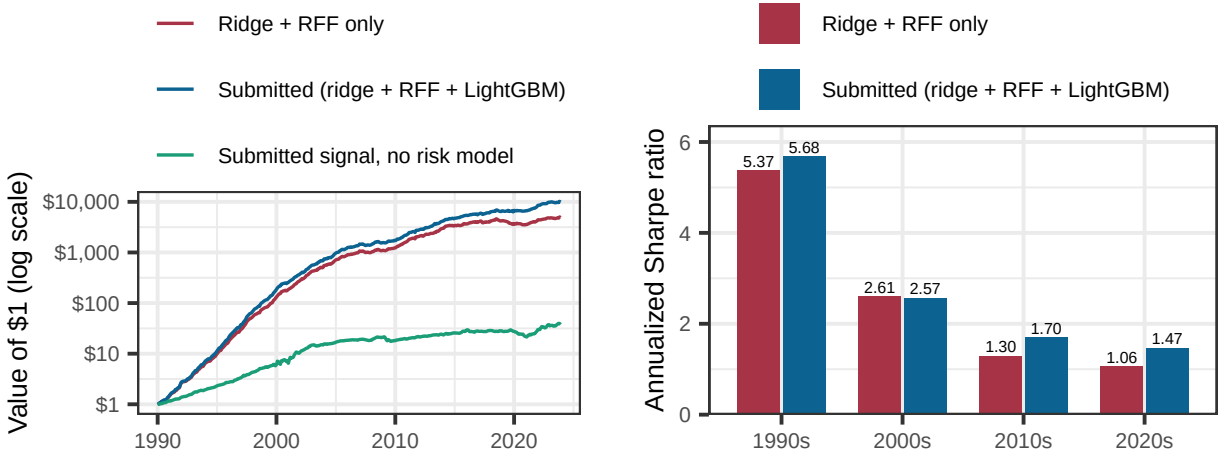


Figure 1. Test period, all with volatility management except the rank portfolio. Left: growth of \$1 invested in December 1989, each strategy scaled to 10% realised volatility; the same signal without the risk model (green) earns far less. Right: Sharpe ratio by decade; adding LightGBM helps most after 2010, when the ridge signals decay the most.

long-short portfolio and 2.79 through the risk model, so most of the Sharpe ratio comes from portfolio construction. The leaderboard is consistent with this: plain ridge, lasso and elastic-net entries score between 0.15 and 0.87, while our ridge-only signal scores 2.53 in our portfolio. LightGBM adds the most in recent years (since 2005 the submitted model scores 1.62 against 1.31 for ridge + RFF), and volatility management makes the risk honest. The top entries pair stronger nonlinear models with more careful covariance estimates, which is where we would push next.

Robustness and What an Investor Should Expect

- **Precision.** With $T = 408$ months, the standard error of the Sharpe ratio, $\sqrt{(1 + \frac{1}{2}SR_m^2)/T}$ annualised (assuming independent months), is 0.20. The mean monthly return has a t-statistic of 16.2, and every decade is positive.

- **Not only a small-stock effect.** The test universe has no micro or nano caps. The gross weight splits 25% mega, 35% large and 40% small caps, and each slice is profitable on its own (Sharpe ratios 1.25, 1.67 and 2.54).
- **Not tuned to one setting.** Without volatility management, rerunning with simpler trees (15 leaves, 1,000 stocks per leaf) gives 5.30 in validation and 2.75 in test, and training on a quarter of the stock-months gives 5.26 and 2.75, against 5.24 and 2.79. We did not use these checks to choose the submission.
- **Costs.** The competition scores returns before costs, as do our headline numbers. With a one-way cost c per dollar traded, the net return is $r_{t+1} - c \sum_i |w_{i,t} - w_{i,t-1}|$. Monthly turnover is about 1.01 times gross exposure, so the net Sharpe ratio is 2.49 at 5 basis points, 2.20 at 10 and 1.29 at 25, with breakeven near 46 basis points; with size-based costs (5, 10 and 25 basis points for mega, large and small caps) it is 1.93. A real fund would add a turnover penalty to the optimiser.

Limitations

- **The characteristics carry hindsight.** They come from the published anomaly literature, much of it written during our test period, and anomaly returns shrink after publication. This applies to every entry, but real-time returns would likely have been lower.
- **Leverage.** Median gross exposure is 7.2 times capital, and volatility management varies it over time. The Sharpe ratio is unaffected, but a real fund would face financing limits.
- **One small look-ahead, in validation only.** The 13 style groups in the risk model are estimated once, from characteristics observed before December 1989. Test-period weights are unaffected, but validation-period portfolios use groupings estimated partly from later characteristics; this affects every risk-model row of Table 1 alike.
- **What we saw, and when.** Sunmeet’s model and its test-period results were shared before LightGBM was added; its portfolio rule was fixed before those results were seen (per Sunmeet’s notes), and we did not test the sensitivity of its Fourier-feature bandwidth or risk-model half-lives. The test results of the first team model were known before Antoine’s volatility management was added; it was adopted by rule R2 on validation data, and on the test period it barely changed the Sharpe ratio.

Credits and disclosure. Anika Bokka prepared the data; Sunmeet Gulati built the ridge and Fourier-feature models and the risk model; Antoine Boisard built the theme model, the significance filter and the volatility management; Connor Hipp built the LightGBM model and the combined pipeline. The baseline follows the course handout, the trees follow Gu, Kelly and Xiu (2020), the risk model the Barra approach from class, the themes Jensen, Kelly and Pedersen (2023) and the volatility scaling Moreira and Muir (2017). We used AI tools for coding and editing, as the course allows. Code: `team_model.R`.