

# Walk-Forward Factor Selection with Meta-Selection

*JKP Characteristics Trading Factor (CTF) challenge - methodology document. Long-short US equity strategy, monthly rebalancing, numpy/pandas/scipy only, deterministic, no external data.*

## 1. Overview and data

The strategy trades a broad cross-section of US stocks using the ~400 firm characteristics from the JKP (Jensen, Kelly & Pedersen 2023) dataset, delivered as the official CTF tables (1.37M stock-months, 1951-2023). It builds one 'factor portfolio' per characteristic, then lets past data - never the test period - choose which factors to hold, how to weight them, and even which selection methodology to trust. The central design goal is the removal of BOTH look-ahead bias and configuration-selection bias, so the reported out-of-sample record is honest.

## 2. The temporal-integrity convention

Each stock-month row carries `ret_exc_lead1m`: the stock's excess return realized over the FOLLOWING month (t+1). It is the prediction target, not a feature, and becomes known only at the end of month t+1. Every factor return stored at month t is therefore a t+1 outcome. The whole strategy is built around one rule: a decision made at the end of month t (which factors, which weights, which configuration) may use only information realized by the end of month t - i.e. factor-return rows dated  $\leq t-1$  month, because the row dated t-1 is the last one realized by end of month t. `ret_exc_lead1m` is used in exactly two legitimate ways: (a) to CONSTRUCT each factor's historical return series, and (b) to SCORE realized performance after the fact. It is never read as a contemporaneous feature or used to make a same-month decision.

## 3. The four stages (and how each avoids look-ahead)

Stage 1 - Factor portfolio returns. For every characteristic and every month, stocks are percentile-ranked cross-sectionally and weighted by  $(rank - 0.5)$ ; the factor's long-short return is  $\sum(w_i \cdot ret_i) / \sum(|w_i|)$  using `ret_exc_lead1m`. This produces a months x factors matrix whose value at month t is realized over t+1. Look-ahead safety: ranks use only month-t characteristics; the return is a pure labelling of the outcome, carried forward untouched.

Stage 2 - Walk-forward simulation of a fixed 96-configuration menu (`window {10y,20y,expanding}` x `n_factors {5,10,15,20}` x `selection {Sharpe,ClusterSharpe}` x `weighting {EW,RankSharpe,MeanVar,InvVol}`). The wider breadth (up to 20 factors) and inverse-volatility (risk-based) weighting are diversification options - chosen a priori, not tuned on the test period - that the meta layer adopts only where past data supports them. Every 6 months each configuration selects and weights factors using ONLY factor-return rows  $\leq$  rebalance minus 1 month, then earns the realized returns until the next rebalance. Look-ahead safety: the selection slice is strictly truncated at  $rp-1$ ; the menu is symmetric and fixed a priori, so no configuration is cherry-picked with hindsight.

Stage 3 - Meta-selection. At each rebalance, each configuration is scored by the annualized Sharpe of its OWN out-of-sample history up to  $rp-1$  (expanding window); the top 3 (once any config has  $\geq 120$  months of track record) are blended equally, else all active configs are blended equally (the EW prior). Look-ahead safety: the methodology live at any date depends only on performance strictly before that date - this is what removes configuration-selection bias.

Stage 4 - Stock weights. For each month the governing rebalance's chosen configurations are mapped to stock weights: stock i receives  $fw_k \cdot (rank_i - 0.5) / \sum_j |rank_j - 0.5|$  summed over factors k, averaged across the chosen configs. Look-ahead safety: stock ranks use only month-t characteristics, and this mapping exactly replicates the factor-level return the earlier stages selected on.

## 4. Fixed design constants (chosen a priori)

Rebalance = 6 months; meta blends top  $K=3$ ; meta starts after 120 months of config history; a factor needs  $\geq 60$  valid months;  $\geq 10$  stocks per side;  $\geq 24$  months before the first rebalance; 30% covariance shrinkage in mean-variance weighting; inverse-vol weighting uses lookback factor standard deviations; ClusterSharpe uses  $\max(3 \cdot N, 20)$  clusters. On small datasets ( $< 200$  months) the observation thresholds scale down so the EW-prior fallback still produces output. No randomness anywhere.

## 5. Results (gross, unit gross leverage)

| Period | Annualized Sharpe | Months |
|--------|-------------------|--------|
|--------|-------------------|--------|

|                           |       |     |
|---------------------------|-------|-----|
| 1960-1989                 | 2.492 | 360 |
| 1990-2003                 | 3.132 | 168 |
| 2004-2013                 | 0.757 | 120 |
| 2014-2023                 | 0.773 | 119 |
| 1990-2023 (out-of-sample) | 1.726 | 407 |

Average monthly one-way turnover: 0.160 of the gross book. Annualized volatility of gross monthly returns: 3.2% at unit gross leverage (1.2% on the book's realized  $\sim 0.39\times$  gross exposure; net exposure  $\sim 0$ ). The post-2004 'honest decades' (Sharpe  $\sim 0.65$ ) are markedly lower than the 1951-2003 sample - exactly the humility the meta-selection design is meant to enforce, and evidence that the numbers are not inflated by hindsight. Returns are gross; a simple net adjustment is  $\text{net}_t = \text{gross}_t - 2 \times \text{turnover}_t \times \text{cost\_per\_side}$ .

## 6. Verification (temporal integrity and determinism)

V1 shift test - PASS. Shifting the entire factor-return matrix one month later leaves the 1990-2023 Sharpe essentially unchanged (1.73  $\rightarrow$  1.72): it neither collapses nor explodes, confirming no stage reads a contemporaneous or future factor return.

V2 truncation test - PASS. Running on data truncated at 1990-01 reproduced all 71 pre-1989 rebalance decisions with zero mismatches against the full-sample run. This is the check the CTF grader automates: re-running on truncated data must reproduce every pre-cutoff decision.

V3 replication identity - PASS. Across 840 months the stock-level portfolio return equals the equally-blended factor-level return of the chosen configurations to 7e-18 (correlation 1.000000), so Stage 4 exactly replicates what Stages 2-3 selected on.

V4 sanity ranges - PASS. Annualized gross volatility is 3.2% at unit gross leverage (1.2% on the book's realized  $\sim 0.39\times$  gross exposure; net exposure  $\sim 0$ ), so at unit gross leverage it sits inside the expected 2-6% band, and no long-period Sharpe exceeds 4 (which would signal  $\text{ret\_exc\_lead1m}$  leaking in as a feature).

V5 determinism - PASS. Two consecutive runs produce identical factor matrix, meta-choices and stock weights; there is no randomness in the pipeline.

## 7. Reproduction

Dependencies are pinned in requirements.txt (numpy 2.5.1, pandas 3.0.3, scipy 1.18.0, pyarrow 25.0.0). The model exposes `main(chars, features, daily_ret) -> DataFrame[id, eom, w]` with no missing weights, and runs start-to-finish with no manual steps. weights.csv contains the portfolio weights for every stock-month; the grader scores only rows flagged `ctff_test`, but all rows are exported.