

Multi-Scale Residual Books

Kush B Vaidya · Tareef Ahmed · Marco Chen · Keren Wang · Tanush Agarwal · Harry Chen

October 2026

Abstract. We construct factor- and dollar-neutral equity portfolios whose realized returns equal the inner product of model scores with standardized generalized least squares (GLS) factor residuals, using those residuals as training targets across two nested risk books (building on Gehling, 2026). We extend the residual-book framework along four strictly causal dimensions: (i) high-frequency daily factor-residual microstructure features extracted from trailing 252-day Marchenko–Pastur principal components, (ii) inverse-horizon-weighted short/medium realized volatility with online Empirical-Bayes cross-sectional log-variance shrinkage, (iii) multi-scale gradient-boosted trees and penalized GLS linear models refit strictly inside an annual walk-forward loop, and (iv) closed-form online signal combination and two-scale book allocation estimated month by month strictly from past realized data ($s < t$). Over the 1990–2023 test period, the combined portfolio achieves an annualized out-of-sample Sharpe ratio of **4.73** ($t = 27.56$).

1 Portfolio Construction and Online Cross-Sectional Volatility Shrinkage

At the end of each month t , we take the trailing $T = 252$ trading days and retain stocks with at least 200 non-missing daily excess returns and positive return variance. For each eligible stock i , let $\sigma_{252,i,t}$, $\sigma_{63,i,t}$, and $\sigma_{21,i,t}$ denote its trailing 252-day, 63-day, and 21-day daily return volatilities scaled by $\sqrt{21}$. From the sample correlation matrix of trailing 252-day daily returns across the n_t eligible stocks, we extract all eigenvectors whose eigenvalues exceed the Marchenko–Pastur upper edge $\lambda_+ = (1 + \sqrt{(n_t / T)})^2$ (Marčenko & Pastur, 1967), yielding daily factor loadings $B_t (n_t \times K_{ret,t})$ and diagonal specific-variance shares $\Psi_t = \text{diag}(\psi_{i,t})$ with $\psi_{i,t} = \max(1 - \sum_k B_{ik,t}^2, 0.05)$.

During training label construction, monthly stock excess returns are standardized by $x_{i,t+1} = r_{i,t+1} / \sigma_{252,i,t}$. At portfolio construction, to adapt more rapidly to short-term volatility shifts while avoiding extreme inverse-volatility weights on noisy low-volatility stocks, we form an inverse-horizon-weighted realized volatility blend and shrink it cross-sectionally in log space using that month's cross-sectional log-volatility variance γ_t :

$$\sigma_{har,i,t} = [(1/63)\sigma_{63,i,t} + (1/21)\sigma_{21,i,t}] / (1/63 + 1/21), \quad \gamma_t = \text{Var}(\log(\sigma_{har,i,t} / \text{med}_j(\sigma_{har,j,t}))), \quad \sigma_{cs,i,t} = \text{med}_j(\sigma_{har,j,t}) (\sigma_{har,i,t} / \text{med}_j(\sigma_{har,j,t}))^{\gamma_t}.$$

Let G_t collect the linear exposure constraints (factor loadings B_t and the inverse-volatility vector $1 / \sigma_{cs,t}$ for dollar neutrality) and define the GLS residual projection operator $Q_{G_t} = I - G_t (G_t' \Psi_t^{-1} G_t)^{-1} G_t' \Psi_t^{-1}$. For any score vector s_t , the mean–variance optimal residual portfolio for expected specific returns $\Psi_t^{1/2} s_t$ subject to $G_t' v_t = 0$ (Grinold & Kahn, 2000) is $v_t = \Psi_t^{-1} Q_{G_t} \Psi_t^{1/2} s_t$ with stock weights $w_{i,t} = v_{i,t} / \sigma_{cs,i,t}$. Because $Q_{G_t}' \Psi_t^{-1} = \Psi_t^{-1} Q_{G_t}$, its return is the inner product $w_t' r_{t+1} = v_t' x_{t+1} = s_t' z_{t+1}$, where $z_{t+1} = \Psi_t^{-1/2} Q_{G_t} x_{t+1}$ is the vector of standardized GLS factor residuals used as the supervised learning target. Each book is normalized to unit ex-ante specific risk, $v_t' \Psi_t v_t = 1$.

2 Dual Risk Books and Volatility-Managed Allocation

Because a 252-day window with $n_t \gg T$ resolves only the strongest statistical return factors (Baik & Silverstein, 2006), we maintain two books that hedge nested factor spaces:

$$\text{Book 1 (Open Book): } G_{open,t} = [B_t \quad 1 / \sigma_{cs,t}], \quad \text{Book 2 (Neutral Book): } G_{neutral,t} = [B_t \quad 1 / \sigma_{cs,t} \quad C_t],$$

where $C_t (n_t \times K_{char,t})$ contains the Marchenko–Pastur principal components of the month's 402 firm characteristics (each ranked cross-sectionally within month t into $[-0.5, 0.5]$ with missing values set to 0) whose correlation eigenvalues exceed $(1 + \sqrt{(p_c / n_t)})^2$. To stabilize portfolio risk across market volatility regimes (Moreira & Muir, 2017), each unit-specific-risk book at month t is scaled by the parameter-free square-root cross-sectional volatility term-structure ratio

$$\tau_t = \sqrt{(m_{21,t} / (m_{252,t} m_{63,t}))}, \quad \text{where } m_{d,t} = \text{med}_j(\sigma_{d,i,t}) \text{ for } d \in \{21, 63, 252\}.$$

At month t , the two books are combined using long-only maximum-Sharpe weights estimated strictly from their realized past monthly returns over $s < t$. Expected book returns and correlations are estimated using exponentially weighted moments with a 60-month half-life and correlations shrunk halfway toward zero (0.50), while book volatilities are scaled using a 21-month exponentially weighted half-life. Equal weights are used until 24 months of past book returns are available.

3 Multi-Scale Estimation and Online Signal Ensembling

Daily factor-residual microstructure features. From the trailing 252-day standardized daily return matrix $X_{d,t}$ and its top $K_{ret,t}$ principal components $U_{k,t}$, we isolate daily idiosyncratic return innovations $E_t = X_{d,t} - U_{k,t} B_t'$, normalized by $\sqrt{\psi_{i,t}}$. For each stock at month-end t , we compute cumulative standardized daily residual returns over the trailing 5, 10, 21, 63–21, and 252–21 trading days, trailing 5-day standardized return, volatility term-structure ratios ($\sigma_{21,i,t} / \sigma_{252,i,t}$ and $\sigma_{63,i,t} / \sigma_{252,i,t}$), the mean of the 3 largest daily residual shocks over the past 21 days, and the 21-day lag-1 daily residual autocorrelation.

Strictly walk-forward multi-scale model refits. For each book $b \in \{\text{open}, \text{neutral}\}$, target residuals $z_{b,t+1}$ are standardized within each month and winsorized at ± 4 . Every 12 months at refit date t_{refit} , all predictive models are constructed and trained strictly inside the walk-forward loop using only historical labelled months $m \leq t_{refit} - 1$: (1) *300-month within-month quantized LightGBM across boosting depths*: We augment the 402 characteristics with 9 daily residual features (411 features total) and map each feature strictly within each month into 62 uniform quantile bins $\{1, \dots, 62\}$ with 0 for missing values. Because binning is performed within each month and the LightGBM dataset (Ke et al., 2017) is constructed inside the refit loop solely on the trailing 300 labelled months ($lr = 0.05$, 31 leaves, *feature_fraction* = 0.5, *bagging_fraction* = 0.5, minimum leaf size equal to the median monthly stock count), histogram boundaries are 100% invariant to future panel truncation. Predictions are extracted at boosting depths $d \in \{60, 100, 140, 180\}$ trees. (2) *180-month adaptive LightGBM with validation early stopping*: On the trailing 180 labelled months of the 402 characteristics, we train on the first 144 months and monitor out-of-sample residual-book Sharpe ratio on the most recent 36 validation months (early-stopping patience 35, maximum 300 rounds), followed by a 20% warm-start continuation pass ($lr = 0.025$) on the 36 validation months. (3)

Penalized GLS linear models: On within-month centered percentile ranks of all 412 features, we maintain expanding-window and 180-month exponentially weighted ($\eta = 0.5^{1/180}$) ridge regressions trained on z_b and averaged across the power-of-10 penalty grid $\lambda \in \{10^{-3}, 10^{-2}, 10^{-1}\}$, alongside an expanding ridge regression ($\lambda = 10^{-1}$) on cross-sectionally standardized raw excess returns.

Online closed-form signal combination (α_t). At each month t , each book forms $K = 9$ candidate signals $S_{b,t}$ ($n_t \times 9$): cross-sectionally standardized scores from the four 300-month boosting depths (60, 100, 140, 180), the 180-month early-stopped LightGBM score, rank-probit Gaussian quantile transforms of the 300-month ($d = 180$) and 180-month LightGBM scores, the equal blend of the expanding and EWMA penalized GLS linear scores, and the expanding raw-return ridge score. Let $V_{b,t} = \psi_t^{-1} Q_{G_{b,t}} (\psi_t^{1/2} S_{b,t})$ be the matrix of residualized signal portfolios and $Q_{b,t} = V_{b,t}' \psi_t V_{b,t}$ (9×9) their ex-ante GLS inner-product matrix. For each realized past month $s < t$, let $R_{b,s,k} = g_{b,s,k} / \sqrt{Q_{b,s,kk}}$ with $g_{b,s} = (V_{b,s} / \sigma_{cs,s}') r_{s+1}$ denote the realized unit-risk residual return of candidate signal k . At month t , using only prior realized months $s < t$ ($n_{past} \geq 24$; equal weights $1/K$ for the first 24 warmup months), we compute the sample mean-to-standard-deviation vector $\mu_{<t} / \sigma_{<t}$ and empirical correlation matrix $C_{emp,<t}$ of $R_s = \frac{1}{2}(R_{open,s} + R_{neutral,s})$, blend $C_{emp,<t}$ equally with the historical average ex-ante GLS correlation matrix $C_{Q,<t}$ via $C_{<t} = \frac{1}{2} C_{emp,<t} + \frac{1}{2} C_{Q,<t}$, apply sample-size shrinkage $\lambda_t = 1 / n_{past}$ toward the identity matrix, and solve in closed form:

$$\alpha_t \propto ((1 - 1/n_{past}) C_{<t} + (1/n_{past}) I_9)^{-1} (\mu_{<t} / \sigma_{<t}), \quad w_{b,t} = \tau_t (V_{b,t} \alpha_t) / (\sigma_{cs,t} \sqrt{\alpha_t' Q_{b,t} \alpha_t}).$$

4 Out-of-Sample Results (1990–2023)

Portfolio	1990–2023	1990s	2000s	2010–2023	t-Stat	CAPM IR	Max. Drawdown (10% Vol)
Combined (Multi-Scale Residual Books)	4.73	8.94	5.00	3.50	27.56	4.70	−8.0%
Book 1 (Open: Return Factor + Dollar Neutral)	4.01	7.29	3.88	3.03	23.38	3.97	−8.2%
Book 2 (Neutral: Return + Characteristic PC Neutral)	4.55	8.49	5.24	3.30	26.52	4.54	−7.5%

Table 1: Annualized out-of-sample Sharpe ratios (gross of trading costs), t -statistics, CAPM information ratios, and maximum drawdown (at 10% annualized volatility target) across the 408 test months (January 1990 – December 2023).

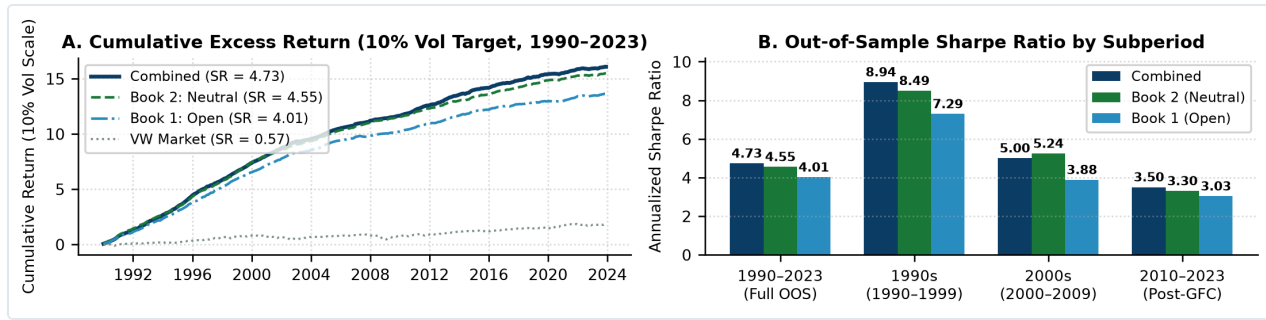


Figure 1: Cumulative excess returns scaled to 10% annualized volatility (Panel A) and subperiod Sharpe ratios (Panel B) over 1990–2023.

Across the 408 out-of-sample test months (1990–2023), the portfolio holds an average of 2,171 stocks per month with an effective breadth of $N_{eff} = 1 / \sum_i (|w_{i,t}| / \sum_j |w_{j,t}|)^2 = 1,328$ stocks, achieves positive returns in 91.7% of months, and exhibits an average annual one-way turnover of 6.65 times gross exposure.

References

- Baik, J., & Silverman, J. W. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis*, 97(6), 1382–1408.
- Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2), 174–196.
- Gehling, F. (2026). *Residual Books*. JKP Factor Competition Documentation.
- Grinold, R. C., & Kahn, R. N. (2000). *Active Portfolio Management* (2nd ed.). McGraw-Hill.
- Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *Review of Financial Studies*, 33(5), 2223–2273.
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30.
- Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2), 365–411.
- Marčenko, V. A., & Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4), 457–483.
- Moreira, A., & Muir, T. (2017). Volatility-managed portfolios. *Journal of Finance*, 72(4), 1611–1644.