

Dynamic Factor Investing with Transformer-Based Return Prediction

Lin Boyi 123090327

Qian Linyi 121090452

Yan Tingyu 124090831

The Chinese University of Hong Kong, Shenzhen

November 9, 2025

Abstract

This proposal outlines a quantitative research project that combines dimension-reduced firm characteristics with Transformer-based sequence models to improve monthly cross-sectional return predictions for China’s A-share market. We begin by diagnosing the instability of traditional linear factor models whose assumptions of stationarity and additivity fail in the presence of regime shifts, evolving regulation, and heterogeneous factor premia. To address these limitations, we propose an empirical pipeline that first applies rolling principal component analysis (PCA) to denoise and compress hundreds of firm-level attributes, followed by a Transformer encoder that ingests 12-month sequences of factor scores to forecast next-month excess returns for CSI 500 constituents. The Transformer will be benchmarked against Ridge regression, Random Forest, and multi-layer perceptron baselines under a rolling-window back-testing protocol with monthly portfolio rebalancing. Evaluation emphasizes information coefficient (IC), IC information ratio (ICIR), and the realized Sharpe ratio from decile-spread portfolios. Drawing on recent advances in machine learning for asset pricing, dynamic factor extraction, and attention-based forecasting, the project aims to demonstrate that merging PCA with attention mechanisms yields more adaptive and robust factor representations than static linear models.

1 Problem Statement

Canonical factor models such as the Fama–French three- or five-factor frameworks and the Hou, Xue, and Zhang (HXZ) q-factor model have become foundational tools for risk decomposition and performance attribution. Yet their usefulness for forward-looking portfolio construction hinges on strong assumptions of linearity and temporal stability. In practice, the relationships between firm fundamentals, exposures, and returns evolve through macroeconomic cycles, policy

interventions, and market microstructure changes. These dynamics are especially pronounced in the Chinese A-share market, where structural breaks emerge from regulatory reforms, shifting investor composition, and state-owned enterprise reforms. When factor loadings shift across regimes, static linear models fail to capture codependencies among characteristics or to adapt to time-varying premia, resulting in diminished predictive power and unstable factor returns.

Moreover, traditional estimators assume that factor returns are generated from a stationary linear combination of observable characteristics. This assumption obscures nonlinear interactions, threshold effects, and higher-order dependencies that have been documented in recent empirical asset pricing studies. For example, the joint effect of profitability and investment may depend on investor sentiment, while liquidity shocks can abruptly alter the performance of style portfolios. The high dimensionality of modern characteristic datasets exacerbates noise and multicollinearity, making it difficult to isolate true signals through ordinary least squares or cross-sectional regressions. As a consequence, factors that appear significant in-sample often decay out-of-sample, leading to model instability and frequent re-specification.

These issues motivate a methodology that can learn flexible, nonlinear mappings between firm characteristics and expected returns while explicitly modeling temporal dependencies. By coupling a denoising stage that produces orthogonal factor scores with a sequence model capable of tracking evolving factor exposures, we aim to move beyond the rigidity of linear factor models and build a more resilient alpha forecasting framework tailored to the regime-sensitive A-share environment.

2 Literature Review

Recent advances in empirical asset pricing underscore the limitations of static linear factor models and motivate the integration of machine learning and dynamic factor extraction. Gu et al.

(2020) demonstrate that carefully tuned machine learning algorithms uncover nonlinear interactions among a high-dimensional set of firm characteristics, delivering superior out-of-sample performance relative to traditional linear benchmarks. Their emphasis on rigorous cross-validation and robust evaluation motivates our inclusion of Ridge regression, Random Forest, and multi-layer perceptron baselines alongside the proposed Transformer architecture. Complementing this perspective, Lettau & Pelger (2020) show that latent factors estimated jointly from the cross-section and time series adapt to shifting market conditions, highlighting the need for models that accommodate time-varying exposures. Their findings inspire our rolling PCA procedure, which refreshes factor loadings each window to reflect evolving A-share dynamics. Building on these insights, Zhang et al. (2023) introduce Finformer, a Transformer-based forecasting framework that captures long-range dependencies and nonlinear temporal patterns more effectively than recurrent networks, providing empirical evidence that attention mechanisms are well suited for financial time series. Together, these studies support a research design that denoises characteristics through PCA while leveraging Transformer attention to model regime-dependent factor structures.

3 Goal and Objectives

The overarching goal of the project is to improve the predictive accuracy and economic value of equity factor investing in the CSI 500 universe through a hybrid machine learning workflow. We articulate the following concrete objectives:

1. Construct a clean panel of monthly firm characteristics and excess returns for CSI 500 constituents, incorporating winsorization, z-score normalization, and optional size- and industry-neutralization to control for pervasive risk exposures.
2. Implement rolling PCA within each training window to extract orthogonal factor scores that

reduce dimensionality while preserving cross-sectional variation in firm fundamentals.

3. Train a Transformer encoder with two layers and four attention heads to map 12-month sequences of PCA factor scores to next-month cross-sectional returns, using mean squared error (MSE) loss and mini-batch optimization.
4. Benchmark the Transformer against Ridge regression, Random Forest, and multi-layer perceptron (MLP) models trained on identical rolling windows and PCA-transformed features.
5. Evaluate predictive and portfolio performance using rolling out-of-sample tests, monthly rebalancing, and metrics including IC, ICIR, Sharpe ratio of a long-short portfolio, and turnover-adjusted returns.

Achieving these objectives will yield practitioner-relevant evidence on whether attention-based architectures can capture dynamic factor structures better than conventional machine learning baselines.

4 Methodology

4.1 Data Universe and Preprocessing

The empirical analysis will focus on CSI 500 constituents to represent medium-sized A-share firms. Daily prices and monthly returns will be aggregated from public data providers such as Wind, CSMAR, or JoinQuant, complemented by fundamental and market microstructure variables commonly used in Chinese factor studies (valuation ratios, profitability margins, leverage, liquidity proxies, quality indicators, growth metrics, and sentiment measures). We will align all variables on a monthly calendar, forward-fill accounting items to avoid look-ahead bias, and remove stocks with less than 24 months of history.

To mitigate the influence of outliers, each characteristic will be winsorized at the 1st and 99th percentiles within each month. We then apply cross-sectional z-score normalization to place features on comparable scales. As a robustness option, we will neutralize characteristics for size and industry by subtracting industry means within CSRC industry buckets, ensuring alpha is not conflated with broad sector tilts.

4.2 Rolling PCA Feature Construction

For each rebalancing date t , we define an expanding or fixed-length training window (e.g., 60 months) that precedes the prediction month. PCA is fitted using only observations inside this window, and we retain the leading principal components that explain at least 80% of the cumulative variance. These components act as latent factors summarizing co-movements among noisy characteristics. The fitted PCA transformation is then applied to stocks in the validation month $t + 1$ to produce orthogonal factor scores. This rolling implementation guards against information leakage and allows factor loadings to evolve over time.

Next, we construct sequences by stacking the most recent 12 months of PCA scores for each stock. Missing months due to listing changes trigger forward-filling up to three months; otherwise, the observation is excluded. The resulting three-dimensional tensor (stock, time, factor score) forms the input for all predictive models to ensure fair comparisons.

4.3 Model Architecture and Training

Our primary model is a Transformer encoder with two stacked layers, each comprising multi-head self-attention (four heads), layer normalization, and position-wise feed-forward networks with GELU activation. Positional encodings are added to capture temporal ordering within each 12-month sequence. We adopt dropout regularization (rate 0.1) and weight decay to mitigate

overfitting. The model outputs a scalar prediction of next-month excess return for each stock, optimized with MSE loss via AdamW. Early stopping based on validation IC prevents over-training, and hyperparameters such as learning rate, embedding size, and feed-forward dimension will be tuned using nested walk-forward validation.

Baseline models share the same rolling training windows and PCA inputs. Ridge regression serves as a linear benchmark with L_2 regularization. Random Forest captures nonlinear interactions through ensemble decision trees, while the MLP offers a deep learning baseline with fully connected layers and ReLU activations. Hyperparameters for each model will be selected through cross-validated grid search within the training window to ensure equitable treatment.

4.4 Evaluation Protocol

We employ a rolling-window back-testing scheme. For each month from 2013 to 2023 (subject to data availability), we train models on the preceding 60 months, validate on the most recent 12 months within the window, and generate out-of-sample predictions for month $t + 1$. Predicted returns are ranked cross-sectionally to compute IC and ICIR. To gauge economic value, we form equal-weighted long-short portfolios that buy the top decile and short the bottom decile of predicted returns, rebalanced monthly. Portfolio returns are adjusted for assumed trading costs (e.g., 30 basis points per side) and evaluated using annualized Sharpe ratios, drawdowns, and turnover metrics. Statistical significance is assessed via Newey–West adjusted t-statistics.

Robustness analyses will inspect sub-period performance across market regimes (bull, bear, sideways), sector concentration, and sensitivity to the number of retained principal components. We also plan to examine attention weights to interpret which factor sequences drive predictions, relating them back to known style factors where possible.

5 Expected Contribution

We anticipate three primary contributions. First, by applying rolling PCA to a rich set of A-share firm characteristics, we reduce measurement noise and mitigate multicollinearity, producing more stable signals for subsequent modeling. Second, the Transformer architecture is designed to capture nonlinear and regime-dependent relationships embedded in the sequences of factor scores, enabling the model to forecast cross-sectional returns more accurately than static linear or tree-based baselines. Third, through comprehensive rolling out-of-sample evaluation and economically meaningful portfolio tests, we will provide evidence on whether the proposed hybrid approach delivers superior IC, ICIR, and Sharpe ratios relative to widely used machine learning alternatives.

Beyond empirical performance, the project will yield interpretability outputs by examining attention distributions across factors and time steps, offering insight into how dynamic factor structures evolve in the CSI 500 universe. The framework can be adapted to other regional markets or integrated with macroeconomic indicators, contributing to the broader literature on data-driven, adaptive factor investing.

6 Project Timeline

Stage	Task	Date Range
1	Kick-off meeting, finalize scope, align data sources, and begin drafting the Problem Statement and Literature Review sections.	November 10, 2024 – November 11, 2024
2	Data acquisition, cleaning routines, validation of CSI 500 characteristic definitions, and completion of the Problem Statement and Literature Review drafts.	November 12, 2024 – November 17, 2024
3	Implement winsorization, z-score normalization, optional size/industry neutralization, and rolling PCA feature store generation.	November 18, 2024 – November 27, 2024
4	Develop Transformer and baseline models, conduct hyper-parameter tuning, and generate preliminary out-of-sample forecasts.	November 28, 2024 – December 4, 2024
5	Portfolio back-testing, robustness diagnostics, interpretability analysis, and preparation of presentation slides.	December 5, 2024 – December 8, 2024
6	Refine documentation, integrate reviewer feedback, and finalize the written report.	December 9, 2024 – December 15, 2024

References

- Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5), 2223–2273.
- Lettau, M., & Pelger, M. (2020). Factors that fit the time series and the cross-section of asset returns. *The Review of Financial Studies*, 33(5), 2274–2325.
- Zhang, Y., Liu, J., Wan, X., Zhang, S., & Wang, R. (2023). Finformer: Transformer-based financial forecasting. *arXiv preprint arXiv:2302.08431*.