FACTOR DIMENSIONALITY AND THE BIAS–VARIANCE TRADEOFF IN DIFFUSION PORTFOLIO MODELS

Avi Bagchi* **Michael Tesfaye*** **Om Shastri**
University of Pennsylvania
Philadelphia, PA 19104, USA
{aviba,tesfaye,oshastri}@seas.upenn.edu

ABSTRACT

In this paper, we implement and evaluate a conditional diffusion model for asset return prediction and portfolio construction on large-scale equity data. Our method models the full distribution of future returns conditioned on firm characteristics (i.e. factors), using the resulting conditional moments to construct portfolios. We observe a clear bias–variance tradeoff: models conditioned on too few factors underfit and produce overly diversified portfolios, while models conditioned on too many factors overfit, resulting in unstable and highly concentrated allocations with poor out-of-sample performance. Through an ablation over factor dimensionality, we reveal an intermediate number of factors that achieves the best generalization and outperforms baseline portfolio strategies.

Track: Industry & Applications

1 INTRODUCTION

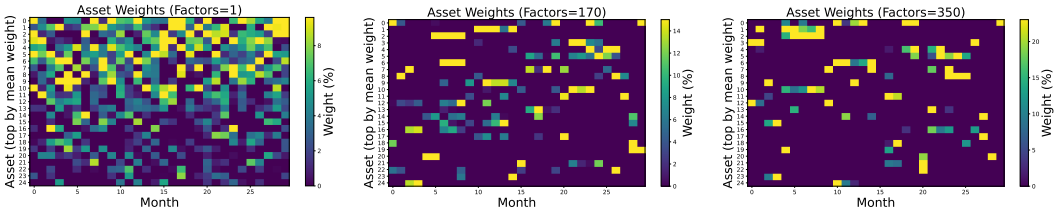
Predicting asset returns is a fundamental problem in quantitative finance. Linear factor models (Fama & French, 1993; 2015) provide a tractable framework for modeling asset returns but struggle to capture nonlinear and higher-order market dynamics. Chen et al. (2026) introduces generative approaches that learn full conditional return distributions rather than point forecasts. In this paper, we evaluate the conditional diffusion framework of Gao et al. (2025), which generates returns conditioned on observable firm characteristics (i.e. factors). We show that factor dimensionality induces a clear bias–variance tradeoff in diffusion-based return modeling: too few factors lead to underfitting and an excessively diverse portfolio, while too many produce high-variance models with overly concentrated allocations. Empirical ablations reveal an optimal dimensionality that outperforms baseline strategies. We use data from Wharton Research Data Services (WRDS) based on the procedure specified by Jensen et al. (2023). We defer dataset details and related work to the appendix (Appendix A.1, Appendix A.2).

2 DIFFUSION-BASED CONDITIONAL RETURN MODELING

We follow Gao et al. (2025) which formulates asset return prediction as learning a conditional return distribution given observable firm characteristics. $R_{t+1} \in \mathbb{R}^N$ denotes a vector of returns for N assets observed in time periods $t = 1 \dots T$ and let $X_t = \{X_{i,j}\}_{i=1}^N$ denote the corresponding set of asset-level characteristics (i.e. factors) observed at time t. Returns are assumed to satisfy $R_{t+1} = f(X_t) + \epsilon_{t+1}$, where $f(\cdot)$ is an unknown, potentially nonlinear function and ϵ_{t+1} captures unpredictable shocks independent of information known at t. The objective is to learn the full conditional distribution $p(R_{t+1}|X_t)$, rather than only conditional means.

To estimate this distribution, we adopt a conditioning denoising diffusion probabilistic model (Ho et al., 2020). The forward diffusion process gradually corrupts observed returns by adding Gaussian noise over a fixed number of steps, transforming the data into an isotropic Gaussian distribution. A neural network is then trained to reverse this process by predicting the noise added at each diffusion

*Equal contribution.



(a) Low capacity (high bias) (b) Medium capacity (c) High capacity (high variance)

Figure 1: Heatmaps for 200 samples show monthly asset weights (top 25 assets by average allocation over the time period) learned under different factor dimensionalities. A low-capacity model (left) distributes weight broadly, reflecting underfitting and high bias. An intermediate model (middle) concentrates allocations on persistent signals, indicating effective factor utilization. A high-capacity model (right) produces sparse, unstable allocations consistent with overfitting and high variance.

step, conditional on characteristics X_t. The reverse diffusion process is implemented using a diffusion transformer architecture. Following Gao et al. (2025) in modifying Peebles & Xie (2023), each asset is represented as a token and cross-sectional dependence among assets is captured through self-attention layers. Conditioning on firm characteristics is performed locally at the token level via adaptive normalization layers. This approach allows the denoising dynamics of each asset to depend on its own characteristics while still modeling joint return behavior across assets. After training, the model generates Monte Carlo samples from the conditional distribution $p(R_{t+1}|x_t)$ for each period, which are used to estimate the conditional mean and covariance of returns that serve as inputs to the portfolio construction procedure (i.e. mean–variance optimization).

3 Results

Each month t, we estimate the conditional mean vector $\hat{\mu}_t$ and the covariance matrix $\hat{\Sigma}_t$ of the next-month returns. We then compute long-only portfolio weights solving a constrained mean-variance optimization problem $\max \quad \omega^\top \hat{\mu}_t - \frac{\gamma}{2}\omega^\top \hat{\Sigma}_t \omega \quad$ subject to $\quad 1^\top \omega = 1 \quad$ and $\quad \omega \geq 0$ with μ as the expected return vector, Σ as the return covariance matrix, $\gamma = 100$ as the risk-aversion parameter, and ω are portfolio weights (Markowitz, 1952). We follow Gao et al. (2025) in comparing the diffusion factor portfolio with three simpler baseline portfolios (Appendix A.3).

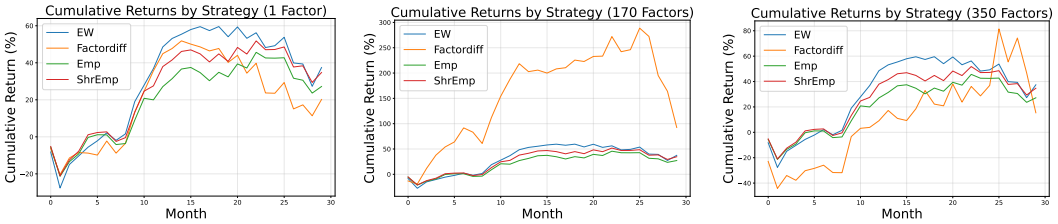
For small k, the portfolio weights are relatively dispersed across assets, reflecting a low-capacity model that produces broadly diversified allocations (Figure 1). As k increases, the weight distribution becomes more concentrated, with larger positions placed on a smaller set of assets (Figure 1). We observe that the moderately diverse portfolio (b) outperforms EW, Emp, and ShrEmp in terms of cumulative returns, where the low capacity (a) ($k = 1$) and the high capacity (c) ($k = 350$) fail to do so (Figure 2). See Appendix B.1 and Appendix B.2 for the full ablation results. Future work should evaluate these results against the framework of Chen et al. (2026; 2023) which implicitly learns a low-dimensional factor structure through score decomposition during score estimation, eliminating the need for explicit factor selection (Appendix B.3).

References

Nicola Borri, Denis Chetverikov, Yukun Liu, and Aleh Tsyvinski. Forward selection fama-macbeth regression with higher-order asset pricing factors, 2025. URL https://arxiv.org/abs/2503.23501.

Mark M Carhart. On persistence in mutual fund performance. *The Journal of Finance*, 52(1):57–82, 1997.

Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data, 2023. URL https://arxiv.org/abs/2302.07194.



(a) Low capacity (high bias) (b) Medium capacity (c) High capacity (high variance)

Figure 2: Cumulative portfolio return for 200 samples over the test months for four strategies. Bias–variance tradeoff illustrated through model capacity. A low-capacity model (left) underfits the data, exhibiting high bias. An intermediate model (middle) achieves a favorable bias–variance balance and the best generalization (we verify this with a larger sample size in Figure 17). An overly expressive model (right) overfits, showing high variance and reduced out-of-sample stability.

Minshuo Chen, Renyuan Xu, Yumin Xu, and Ruixun Zhang. Diffusion factor models: Generating high-dimensional returns with factor structure, 2026. URL https://arxiv.org/abs/2504.06566.

Eugene F Fama and Kenneth R French. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56, 1993.

Eugene F Fama and Kenneth R French. A five-factor asset pricing model. *Journal of Financial Economics*, 116(1):1–22, 2015.

Jianqing Fan, Yingying Fan, and Jinchi Lv. High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147(1):186–197, 2008.

Jianqing Fan, Yuan Liao, and Martina Mincheva. Large covariance estimation by thresholding principal orthogonal complements, 2013. URL https://arxiv.org/abs/1201.0175.

Guanhao Feng, Wei Lan, Hansheng Wang, and Jun Zhang. Selecting and testing asset pricing models: A stepwise approach, 2026. URL https://arxiv.org/abs/2601.10279.

Xuefeng Gao, Mengying He, and Xuedong He. Factor-based conditional diffusion model for portfolio optimization, 2025. URL https://arxiv.org/abs/2509.22088.

Richard C. Grinold and Ronald N. Kahn. *Active Portfolio Management: A Quantitative Approach for Producing Superior Returns and Controlling Risk*. McGraw-Hill, 2 edition, 2000.

Shihao Gu, Bryan Kelly, and Dacheng Xiu. Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273, 2020.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239.

William James and Charles Stein. Estimation with quadratic loss. In Samuel Kotz and Norman L. Johnson (eds.), *Breakthroughs in Statistics: Foundations and Basic Theory*, pp. 443–460. Springer, New York, 1992. Reprint of the original 1961 paper.

Theis Ingerslev Jensen, Bryan Kelly, and Lasse Heje Pedersen. Is there a replication crisis in finance? *The Journal of Finance*, 78(5):2465–2518, 2023.

Chen Jin and Ankush Agarwal. Forecasting implied volatility surface with generative diffusion models, 2025. URL https://arxiv.org/abs/2511.07571.

Bryan T. Kelly, Seth Pruitt, and Yinan Su. Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics*, 134(3):501–524, 2019. doi: 10.1016/j.jfineco.2019.05.010. URL https://www.nber.org/papers/w24540. NBER working paper version available; published version in JFE.

Gihun Kim, Sun-Yong Choi, and Yeoneung Kim. A diffusion-based generative model for financial time series via geometric brownian motion, 2025. URL https://arxiv.org/abs/2507.19003.

Olivier Ledoit and Michael Wolf. Honey, i shrunk the sample covariance matrix. *The Journal of Portfolio Management*, 30(4):110–119, 2004.

Harry Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.

Caspar Meijer and Lydia Y. Chen. The rise of diffusion models in time-series forecasting, 2024. URL https://arxiv.org/abs/2401.03006.

William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023. URL https://arxiv.org/abs/2212.09748.

Kashif Rasul, Calvin Seward, Ingmar Schuster, and Roland Vollgraf. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting, 2021. URL https://arxiv.org/abs/2101.12072.

Barr Rosenberg and Vijay Marathe. Extra-market components of covariance in security returns. *Journal of Financial and Quantitative Analysis*, 1974.

Lifeng Shen and James Kwok. Non-autoregressive conditional diffusion models for time series prediction, 2023. URL https://arxiv.org/abs/2306.05043.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2021. URL https://arxiv.org/abs/2011.13456.

Chen Su, Zhengzhou Cai, Yuanhe Tian, Zhuochao Chang, Zihong Zheng, and Yan Song. Diffusion models for time series forecasting: A survey, 2025. URL https://arxiv.org/abs/2507.14507.

Tomonori Takahashi and Takayuki Mizuno. Generation of synthetic financial time series by diffusion models, 2024. URL https://arxiv.org/abs/2410.18897.

Yuki Tanaka, Ryuji Hashimoto, Takehiro Takayanagi, Zhe Piao, Yuri Murayama, and Kiyoshi Izumi. Cofindiff: Controllable financial diffusion model for time series generation, 2025. URL https://arxiv.org/abs/2503.04164.

Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. Csdi: Conditional score-based diffusion models for probabilistic time series imputation, 2021. URL https://arxiv.org/abs/2107.03502.

Zhuohan Wang and Carmine Ventre. A financial time series denoiser based on diffusion model, 2024. URL https://arxiv.org/abs/2409.02138.

A Appendix

A.1 Data

Our analysis uses the Global Factor Data constructed by Jensen, Kelly, Pederson and distributed through Wharton Research Data Services (WRDS). The dataset combines information from CRSP and Compustat to provide a comprehensive panel of firm-level characteristics and return for publicly traded equities. The dataset includes data from January 2010 until February 2025. The data includes more than 400 characteristics constructed following the procedures documented in (Jensen et al., 2023).

To align the WRDS factor data with the diffusion framework, we apply standard cross-sectional preprocessing and construct a fix-shape monthly panel. We restrict the sample to U.S. common stocks and define the prediction target as next-month return by shifting realized returns forward. Returns are winsorized cross-sectionally to mitigate outliers, while firm characteristics are imputed

using cross-sectional means, standardized, and clipped within each month. For each month, we retain a fixed number of assets and organize the resulting data into tensors of characteristics and returns with dimensions (T, N, K) and (T, N), respectively, where T denotes the number of months, N denotes the number of assets, and K the number of firm-level characteristics. We use $T = 150$, $N = 200$, $K = 350$. These tensors serve as inputs to the conditional diffusion model.

A.2 Related Work

Diffusion Models: Diffusion models learn complex data distributions by progressively corrupting data with noise and training a neural network to reverse this process. The model learns the score function via score matching, enabling sampling by iteratively denoising from noise back to data (Ho et al., 2020; Song et al., 2021). For time series, prior work uses diffusion either (i) as a conditional scenario generator for forecasting or (ii) as a conditional model for missing-data problems (Rasul et al., 2021; Tashiro et al., 2021). Surveys synthesize highlight evaluation pitfalls in diffusion-based time-series forecasting (Meijer & Chen, 2024; Su et al., 2025).

In finance, diffusion is mainly used as a conditional scenario generator for returns, with several papers emphasizing controllability or finance-specific noise structure. Shen et al. propose a non-autoregressive conditional diffusion model for generating future time-series trajectories conditioned on historical data, which can be applied to financial return forecasting (Shen & Kwok, 2023). Tanaka et al. focus on controllable conditional generation for financial time series, adding explicit controls to steer the sampled trajectories toward desired attributes Tanaka et al. (2025). Kim et al. modify the diffusion forward noising process to reflect financial structure (e.g., heteroskedasticity and multiplicative noise), targeting more realistic synthetic dynamics and improved conditional sampling Kim et al. (2025). Wang et al. study finance-tailored denoisers and Takahashi et al. propose methods aimed at synthetic financial time-series generation with finance-specific modeling choices (Wang & Ventre, 2024; Takahashi & Mizuno, 2024). Beyond return/path generation, Jin et al. apply diffusion in an option-centric setting by forecasting the implied-volatility surface (Jin & Agarwal, 2025).

Factor Models Factor models are a standard framework for portfolio construction, with widely used specifications such Fama & French (1993; 2015); Carhart (1997). Modern portfolio risk systems build on this framework by estimating factor exposures and covariance structures Rosenberg & Marathe (1974) and its practical development for quantitative portfolio construction Grinold & Kahn (2000). A limitation of factor models is estimation error in high dimensions (Ledoit & Wolf, 2004; Fan et al., 2008; 2013). Work in empirical asset pricing shows that large panels of firm characteristics improve return prediction but introduce redundancy and model selection challenges (Gu et al., 2020; Kelly et al., 2019). As the number of proposed factors has expanded, recent studies emphasize systematic testing, dimensionality control, and the risk of overfitting in high-dimensional factor spaces (Feng et al., 2026; Borri et al., 2025).

A.3 Portfolio Constructions

In the transaction cost setting, we augment the objective with linear trading costs. Portfolio returns are computed as the inner product of portfolio weights and realized returns, and performance is summarized using mean return, volatility, and annualized Sharpe ratio.

- **Equal-Weighted (EW)** assigns uniform weights and does not estimate return moments.
- **Empirical (Emp)** estimates mean and covariance directly from historical returns using a rolling window.
- **Shrinkage Empirical (ShrEmp)** applies covariances shrinkage to improve stability while retaining the sample covariance mean (James & Stein, 1992).

We obtain moment estimates the conditional distribution of next-period returns using the conditional diffusion model outlined above. Monte Carlo samples drawn from this distribution are used to estimate the conditional mean and covariance of returns. We use 200 samples in the results diagrams.

B Additional Figures

Let k denote the number of factors. We perform an ablation for $k \in \{1, 3, 6, 11, 18, 30, 48, 75, 115, 170, 240, 300, 350\}$.

B.1 Cumulative Returns

The figures in this section illustrate how performance varies as the number of factors k increases. When k is small, the model is overly constrained and fails to capture sufficient structure in the data. This high-bias regime leads to underfitting: cumulative returns closely track or underperform the baseline. As k increases, performance improves and the model begins to capture meaningful relationships. It begins to out-perform the baseline when $k \geq 18$. However, for very large k, performance deteriorates again. The model enters a high-variance regime in which additional factors primarily fit noise rather than signal. This is reflected in reduced out-of-sample performance, with returns again failing to outperform the baseline.

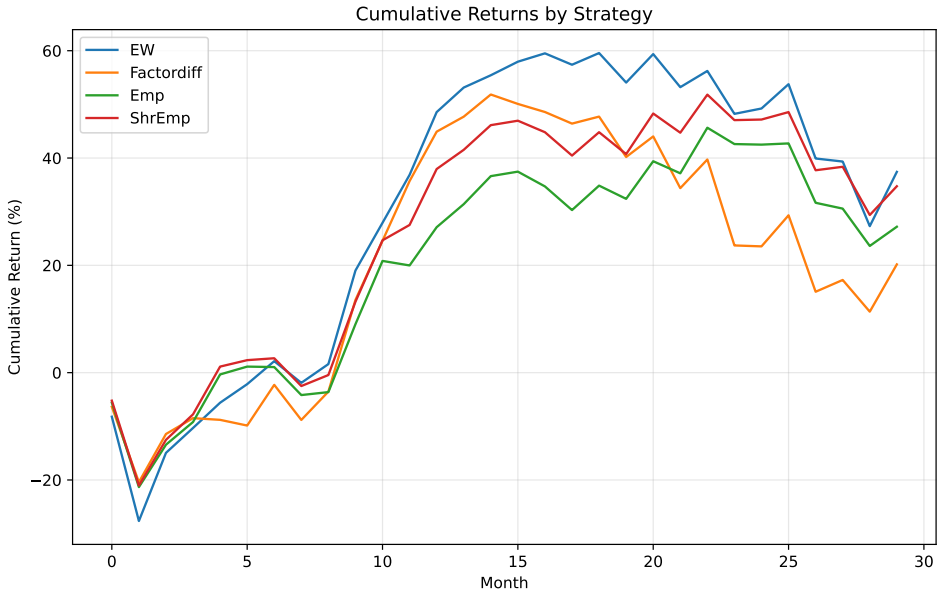


Figure 3: Cumulative returns ($k = 1$)

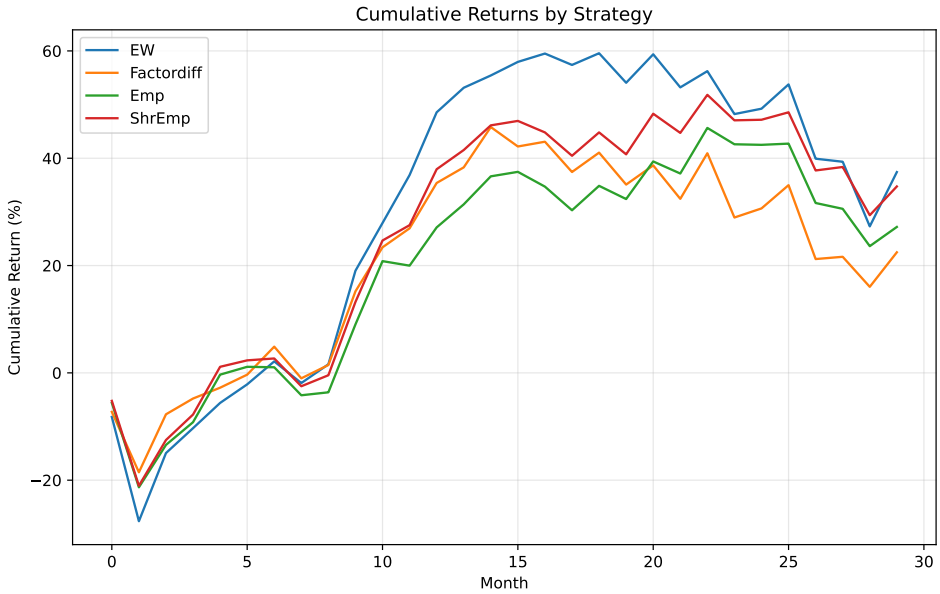


Figure 4: Cumulative returns ($k = 3$)

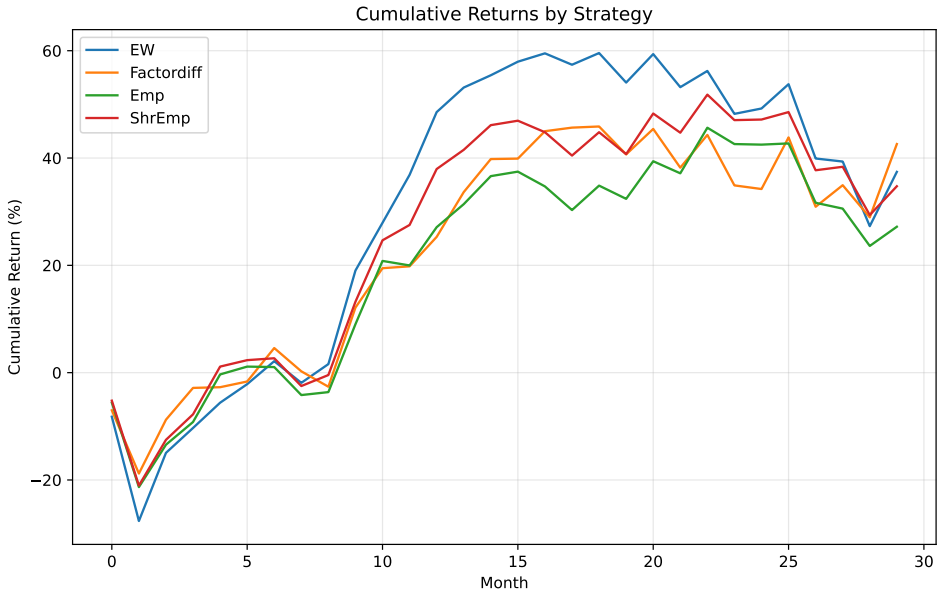


Figure 5: Cumulative returns ($k = 6$)

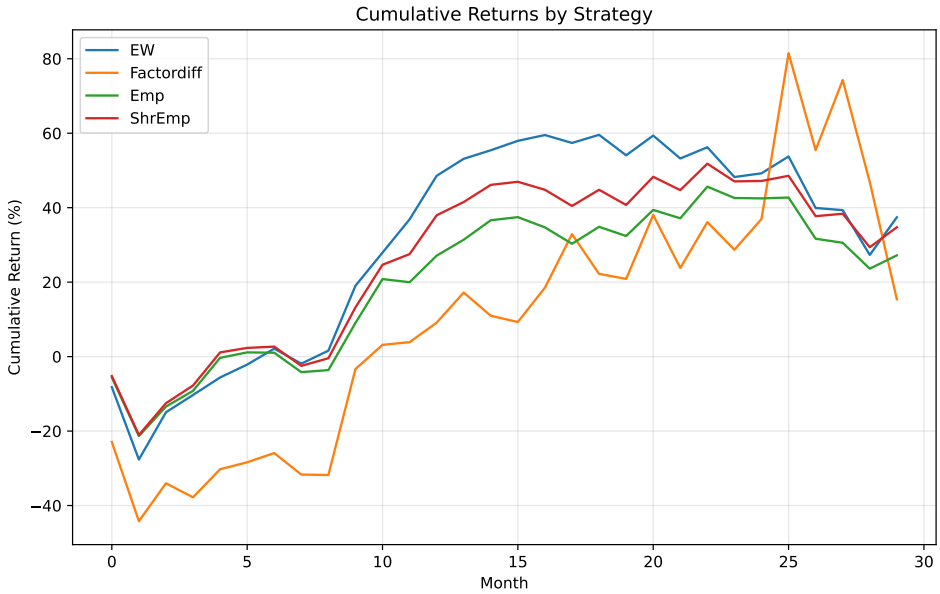


Figure 6: Cumulative returns ($k = 10$)

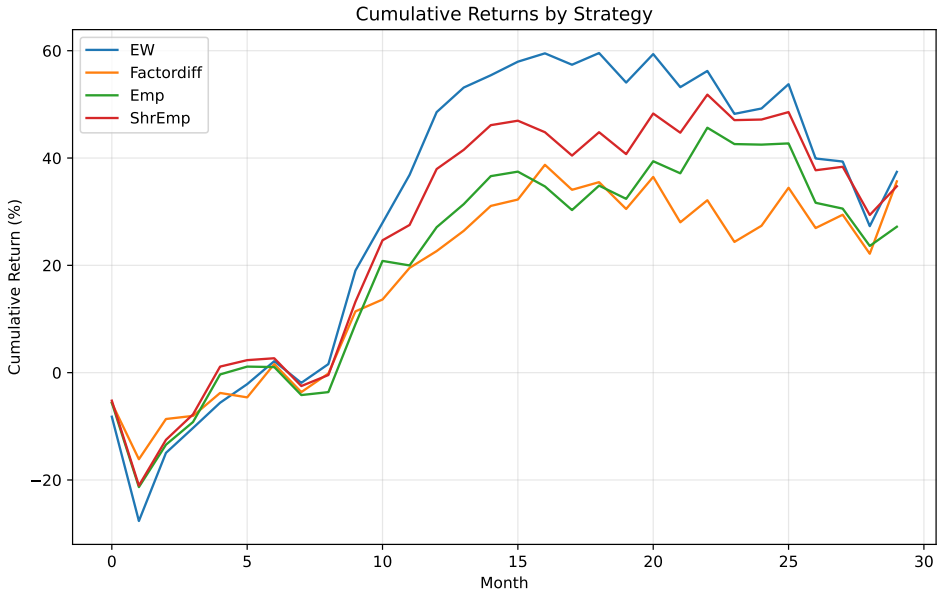


Figure 7: Cumulative returns ($k = 11$)

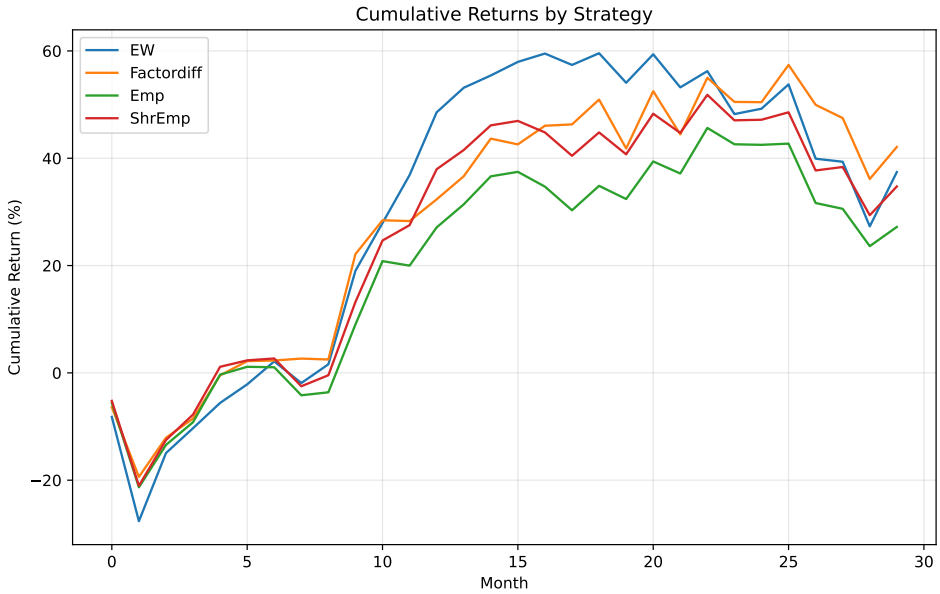


Figure 8: Cumulative returns ($k = 18$)

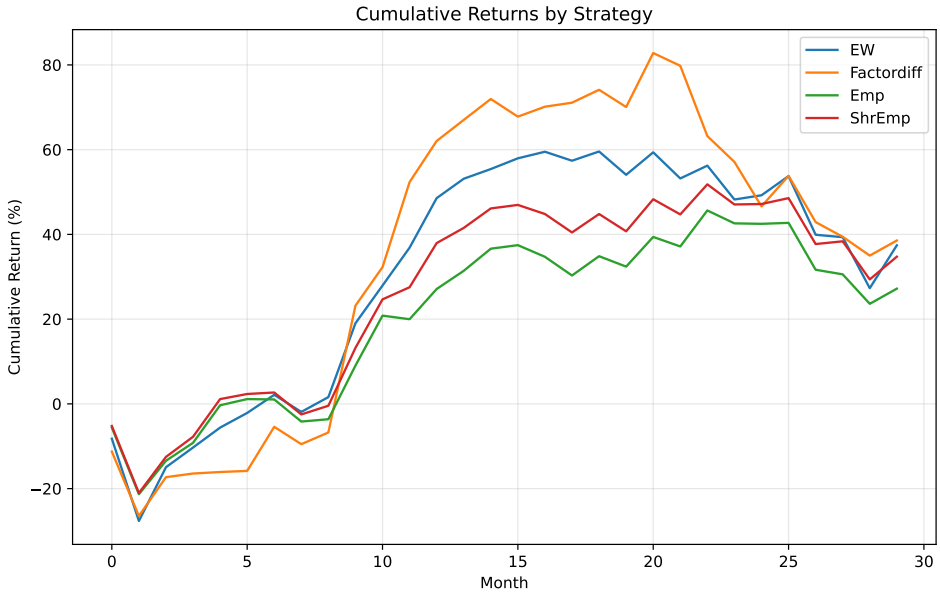


Figure 9: Cumulative returns ($k = 30$)

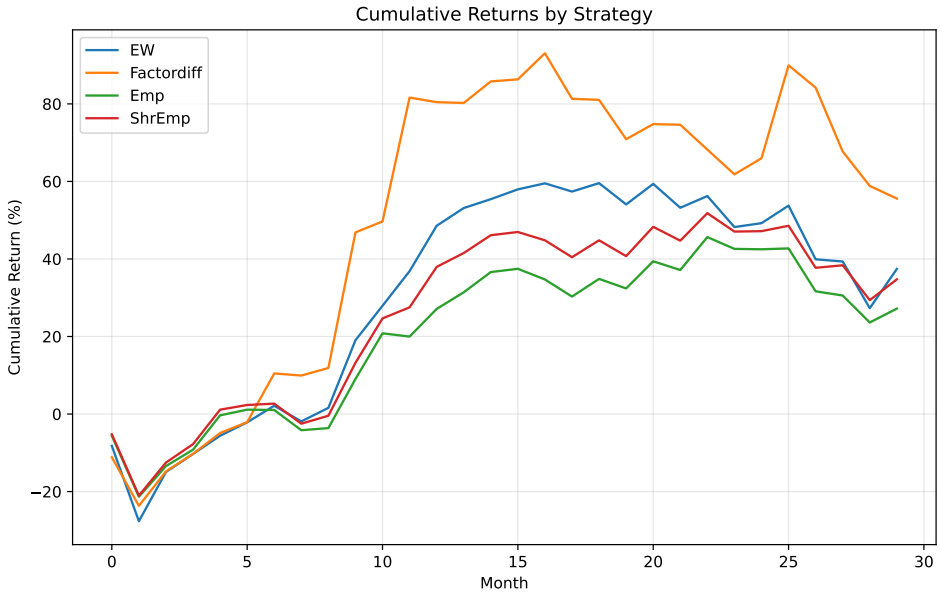


Figure 10: Cumulative returns ($k = 48$)

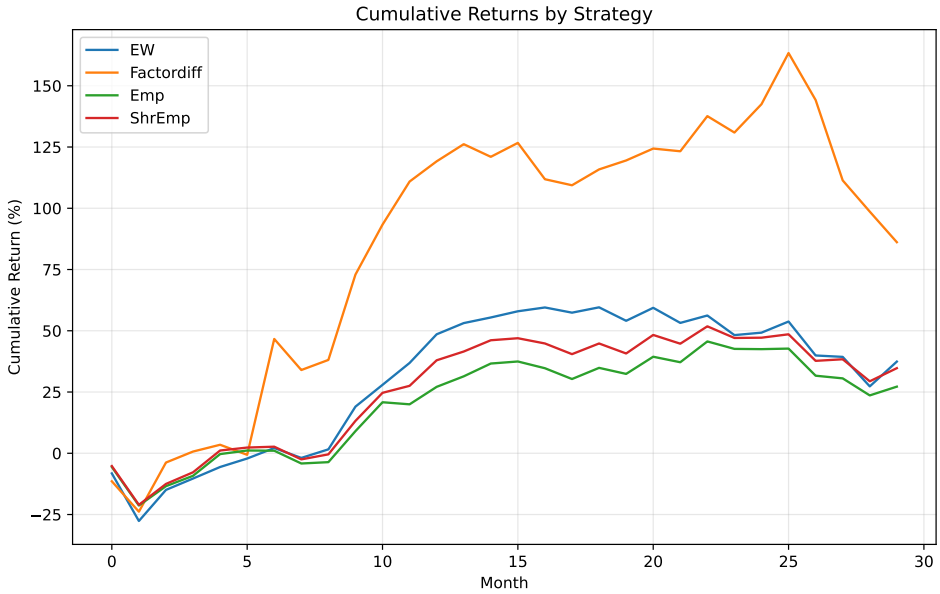


Figure 11: Cumulative returns ($k = 75$)

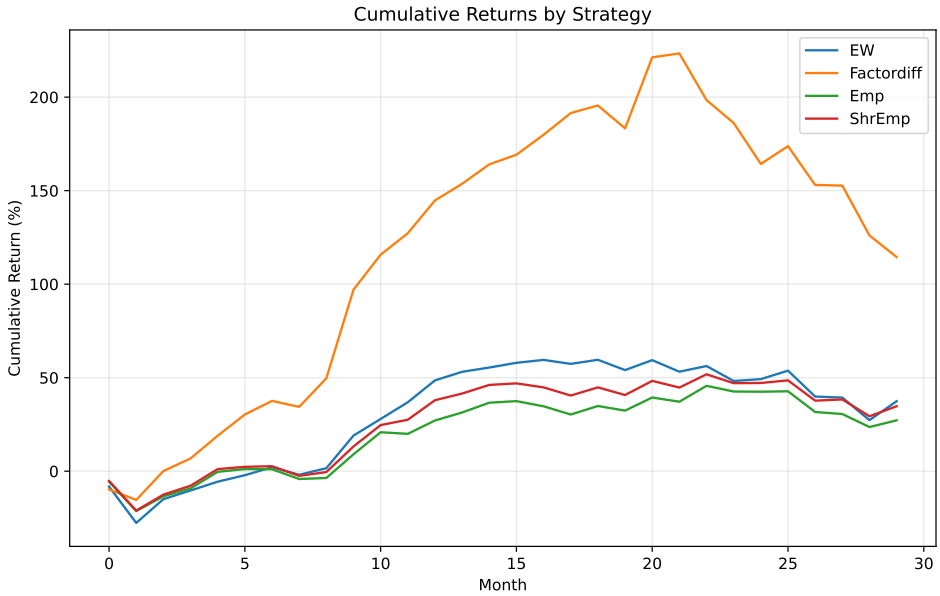


Figure 12: Cumulative returns ($k = 115$)

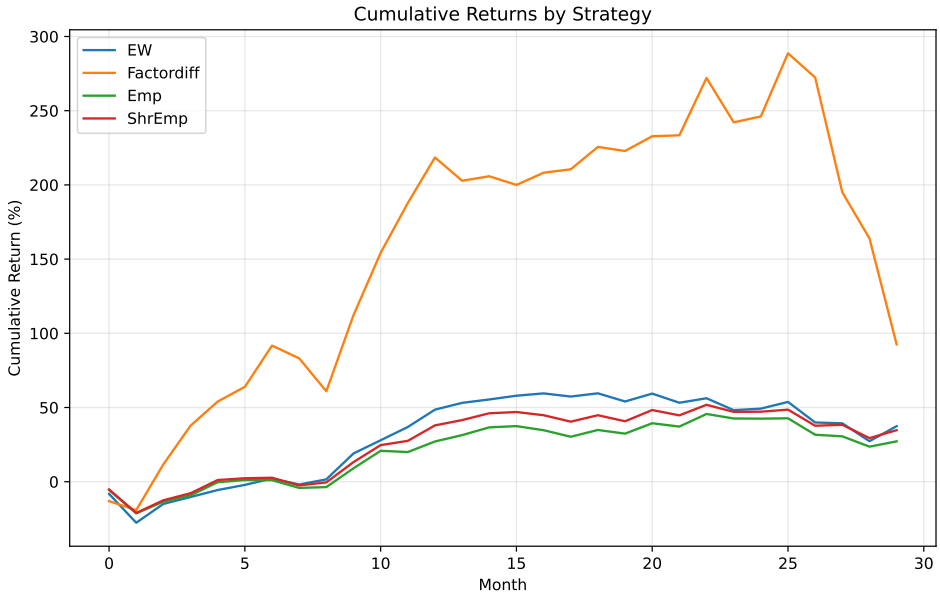


Figure 13: Cumulative returns ($k = 170$)

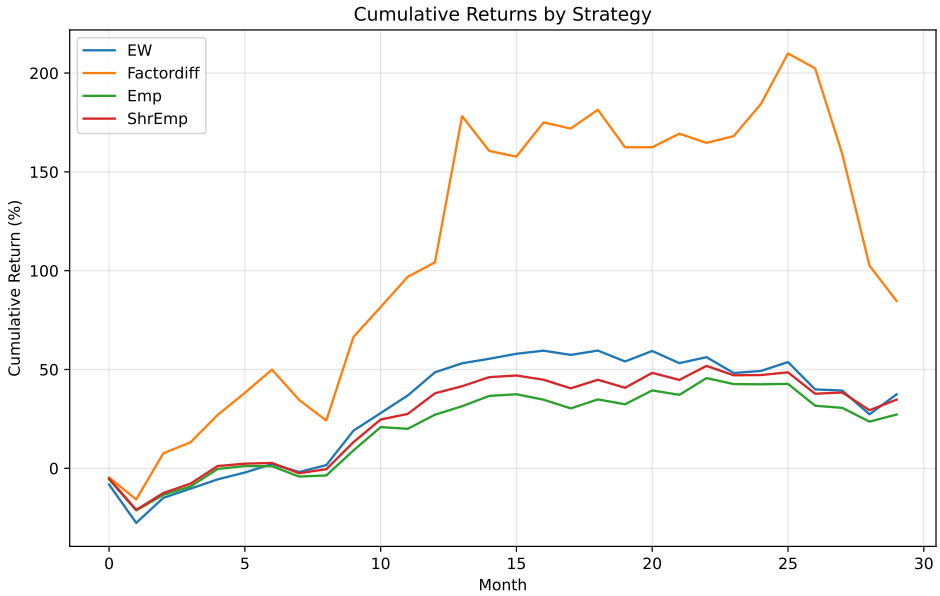


Figure 14: Cumulative returns ($k = 240$)

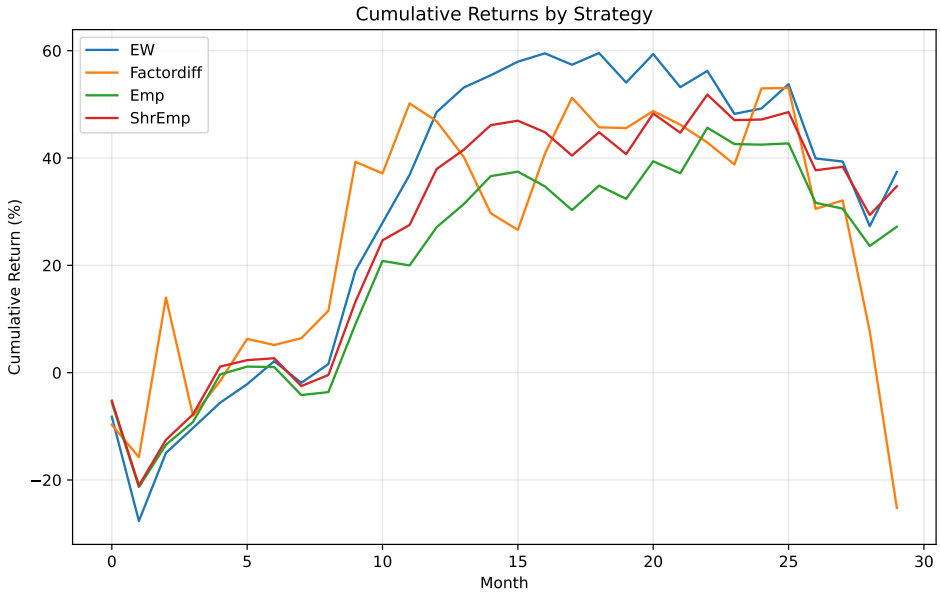


Figure 15: Cumulative returns ($k = 300$)

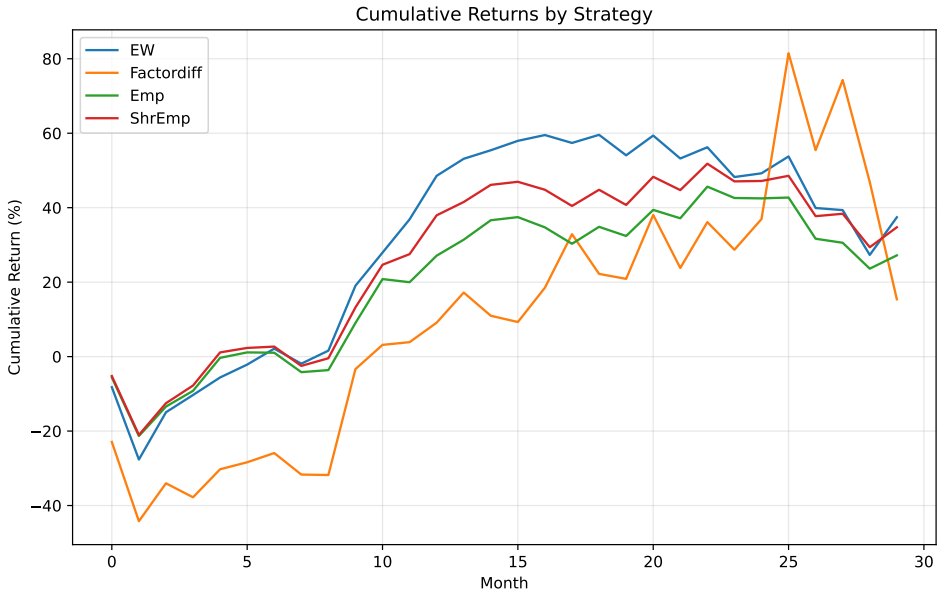


Figure 16: Cumulative returns ($k = 350$)

We verify the performance of $k = 170$, raising the number of samples to 1000, again observing that it outperforms the baseline.

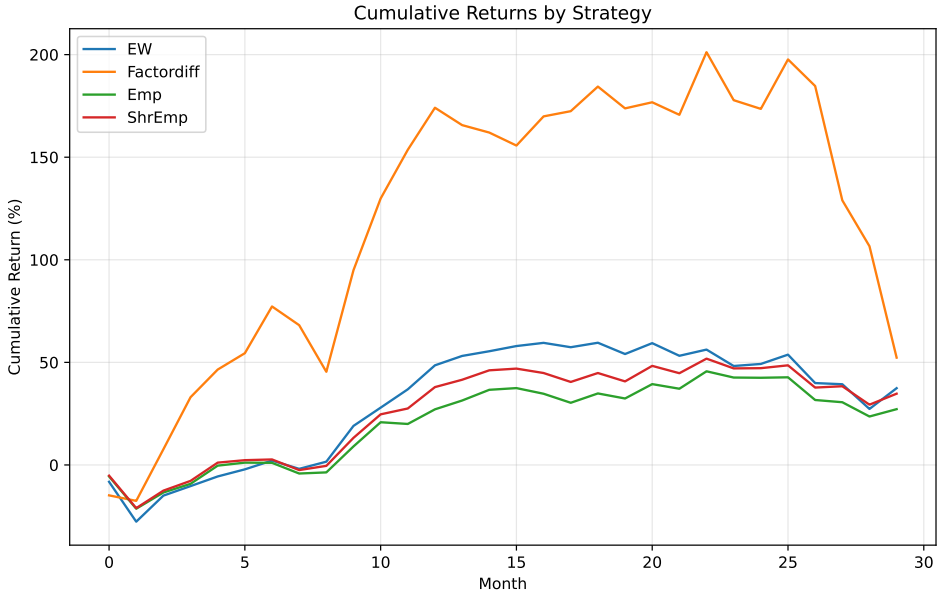


Figure 17: Cumulative returns ($k = 170$)

B.2 PORTFOLIO WEIGHTS

In this section, we show how increasing the number of factors k progressively changes the structure of the learned portfolios. For small k, the weights are broadly distributed across many assets, indicating a diffuse allocation consistent with an under-parameterized model. As k grows, the portfolio weights become increasingly concentrated, with larger magnitudes assigned to a smaller subset of assets. This concentration suggests that higher-capacity models identify more specific signals but also become more sensitive to noise, leading to over-specialized allocations and reduced out-of-sample performance.

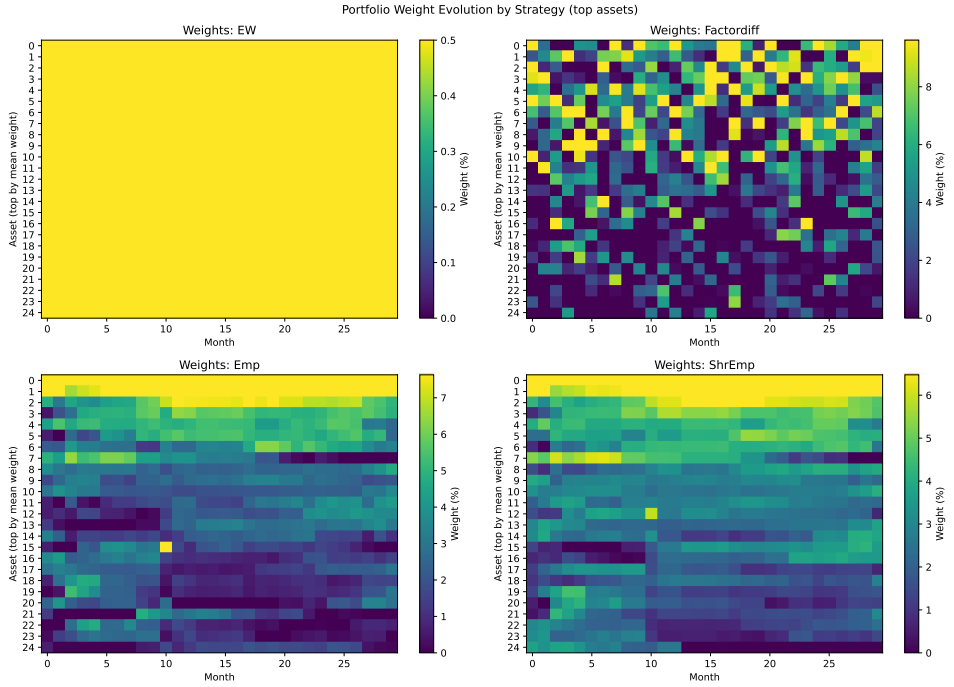


Figure 18: Weights heatmap ($k = 1$)

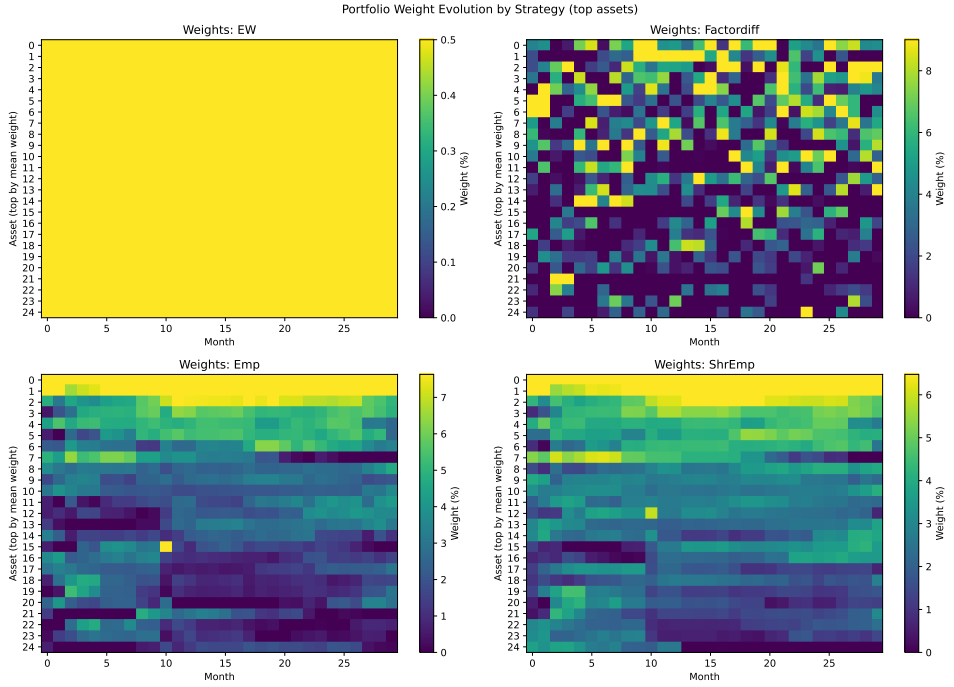


Figure 19: Weights heatmap ($k = 3$)

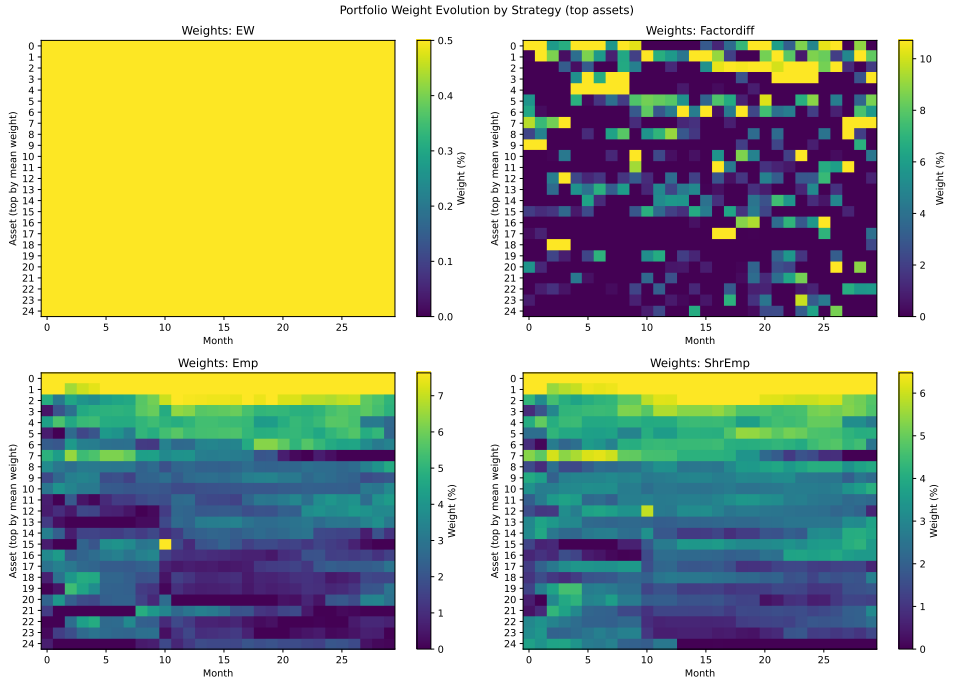


Figure 20: Weights heatmap ($k = 6$)

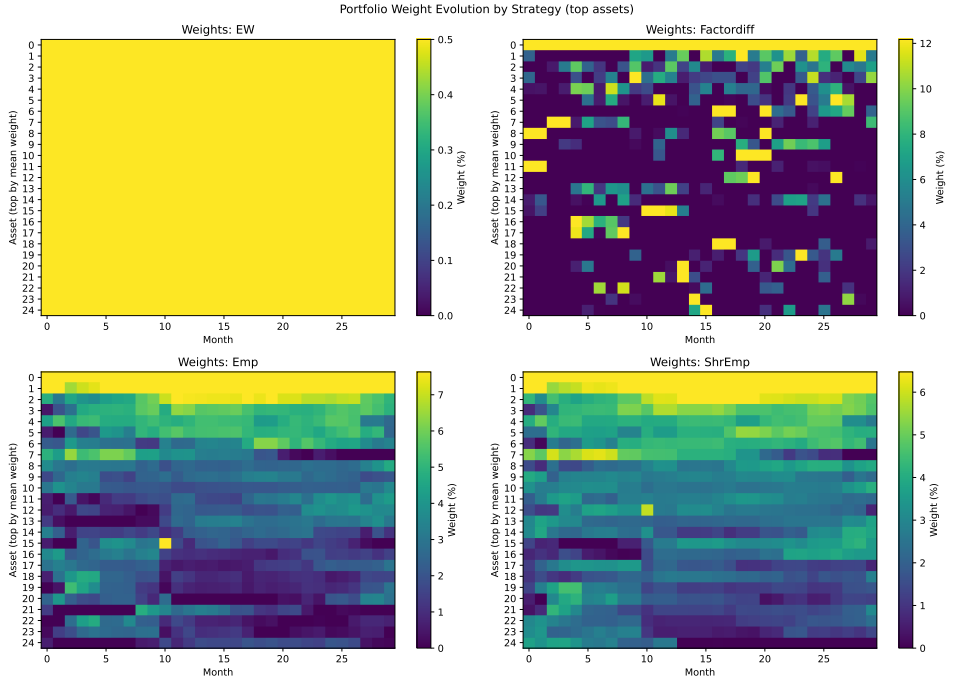


Figure 21: Weights heatmap ($k = 11$)

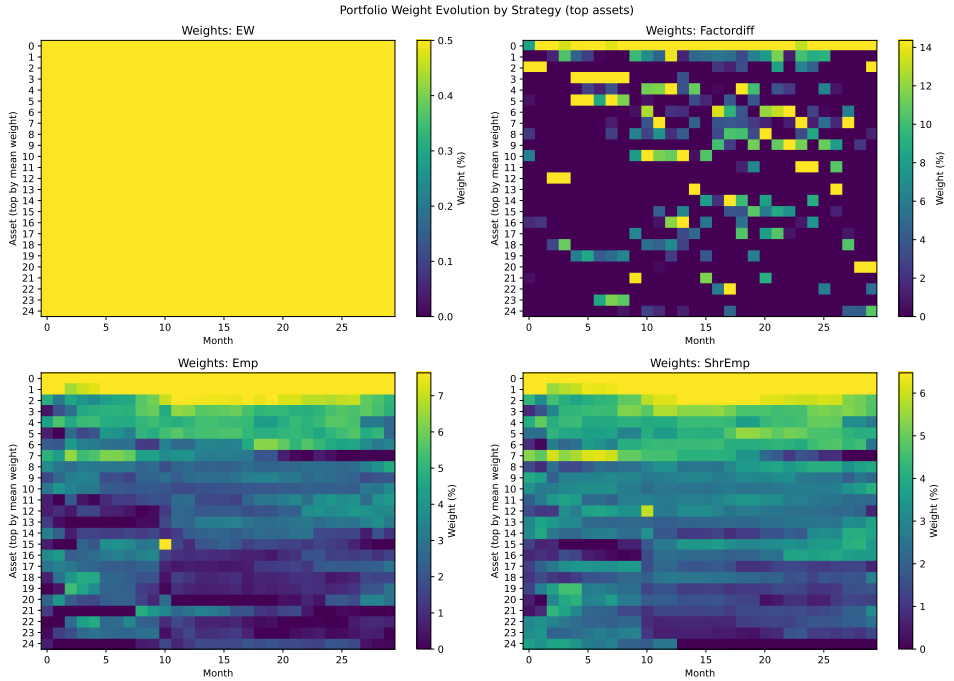


Figure 22: Weights heatmap ($k = 18$)

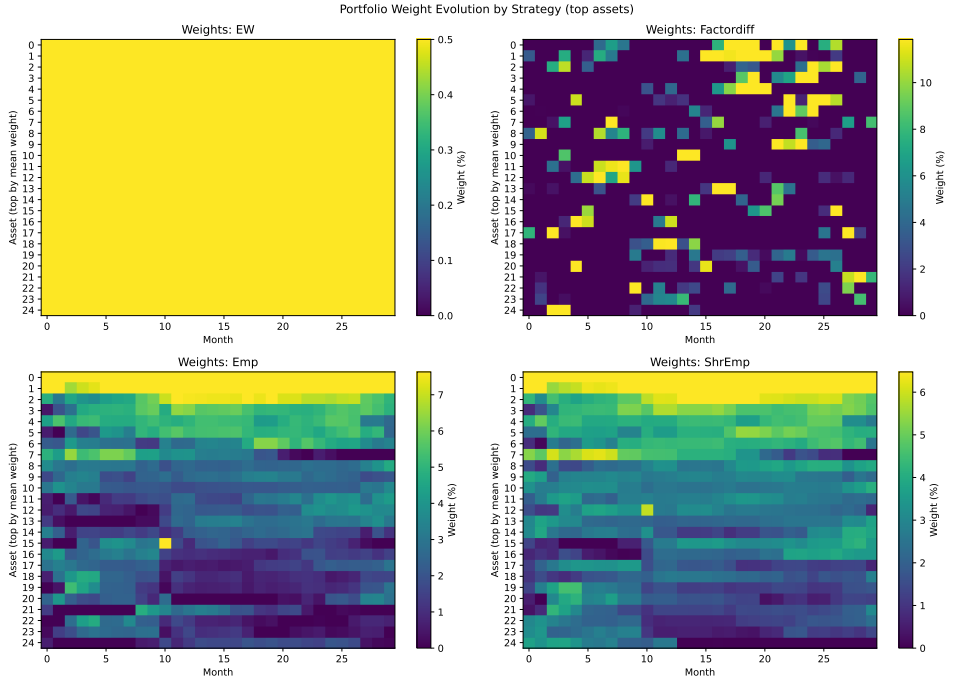


Figure 23: Weights heatmap ($k = 30$)

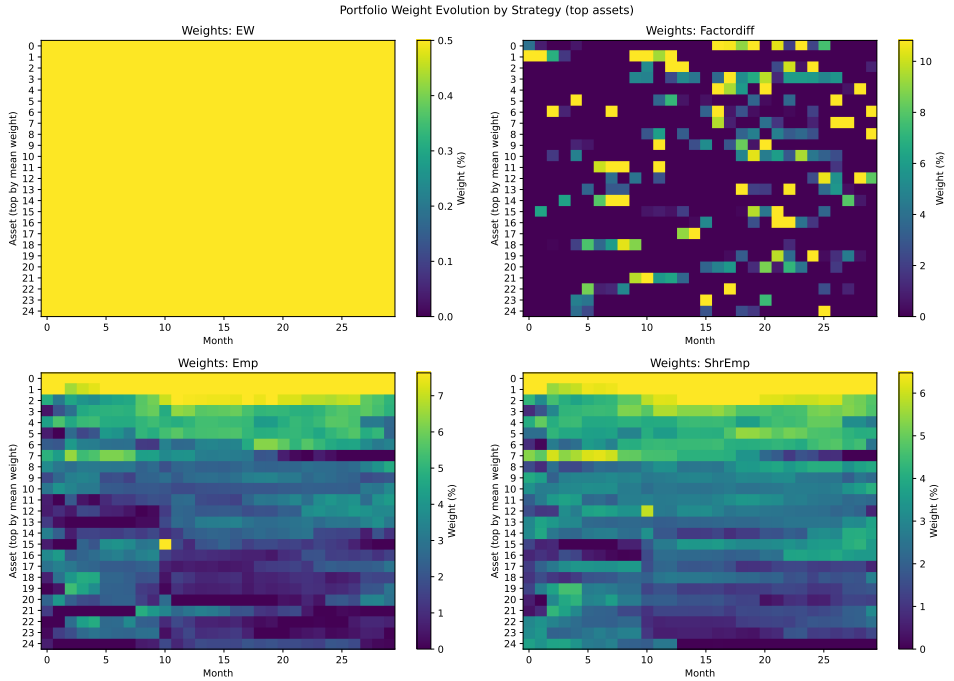


Figure 24: Weights heatmap ($k = 48$)

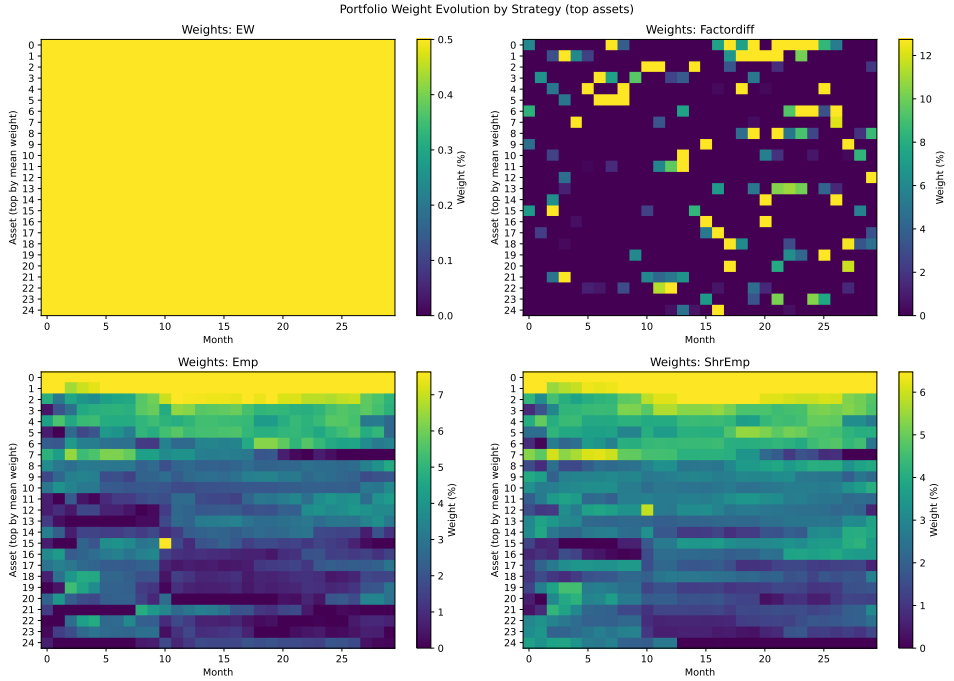


Figure 25: Weights heatmap ($k = 75$)

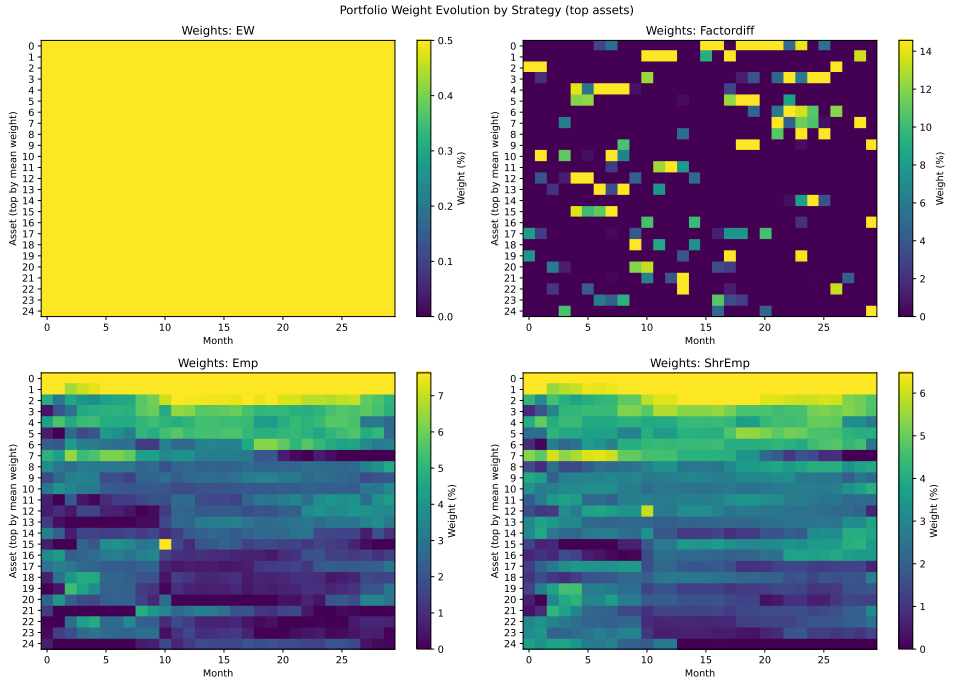


Figure 26: Weights heatmap ($k = 115$)

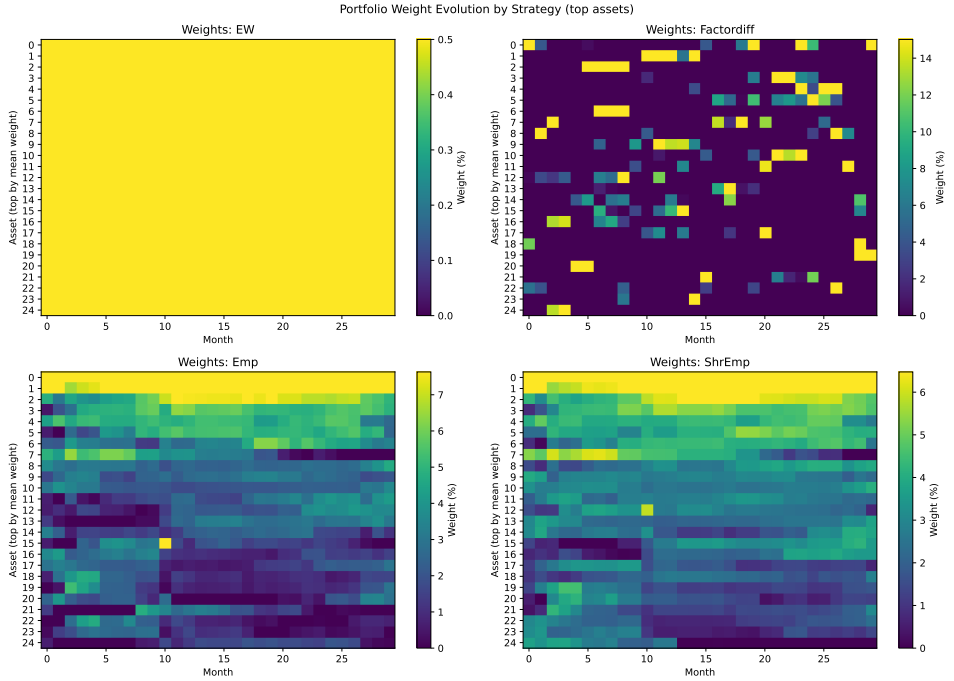


Figure 27: Weights heatmap ($k = 170$)

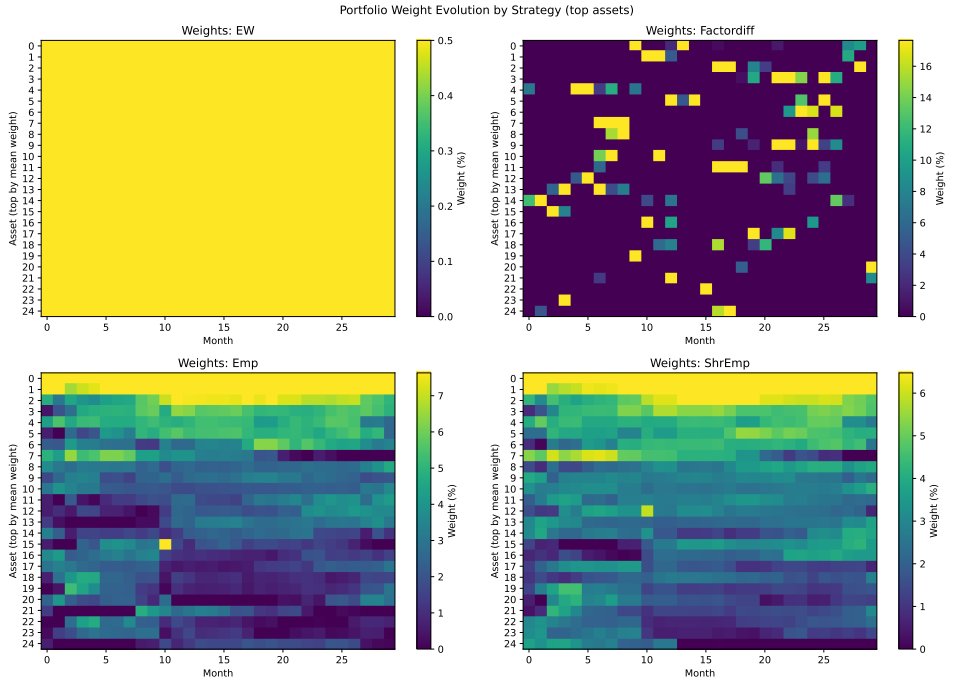


Figure 28: Weights heatmap ($k = 240$)

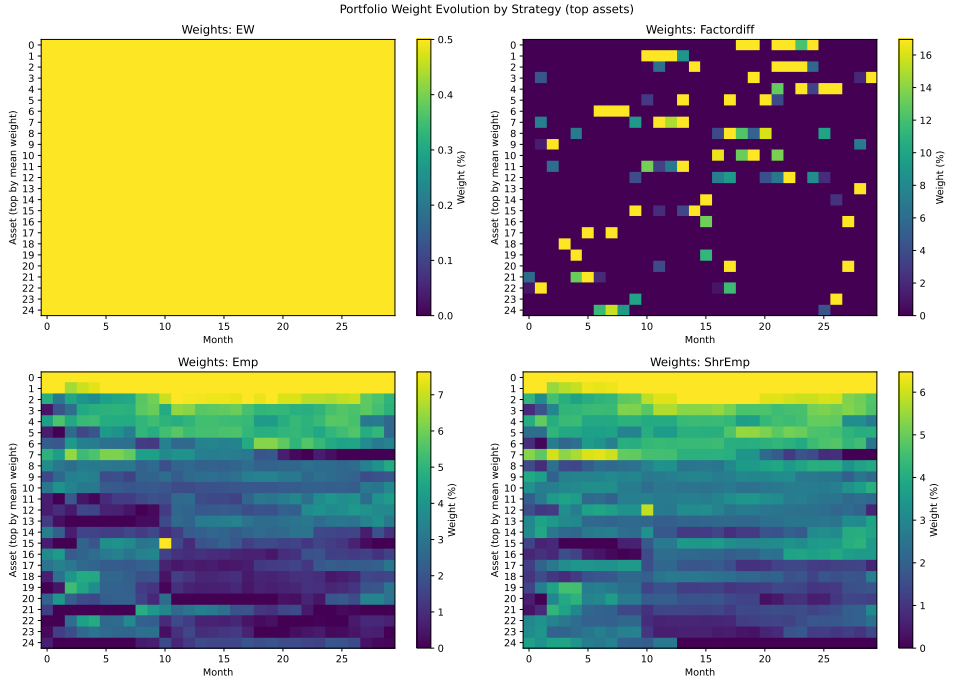


Figure 29: Weights heatmap ($k = 300$)

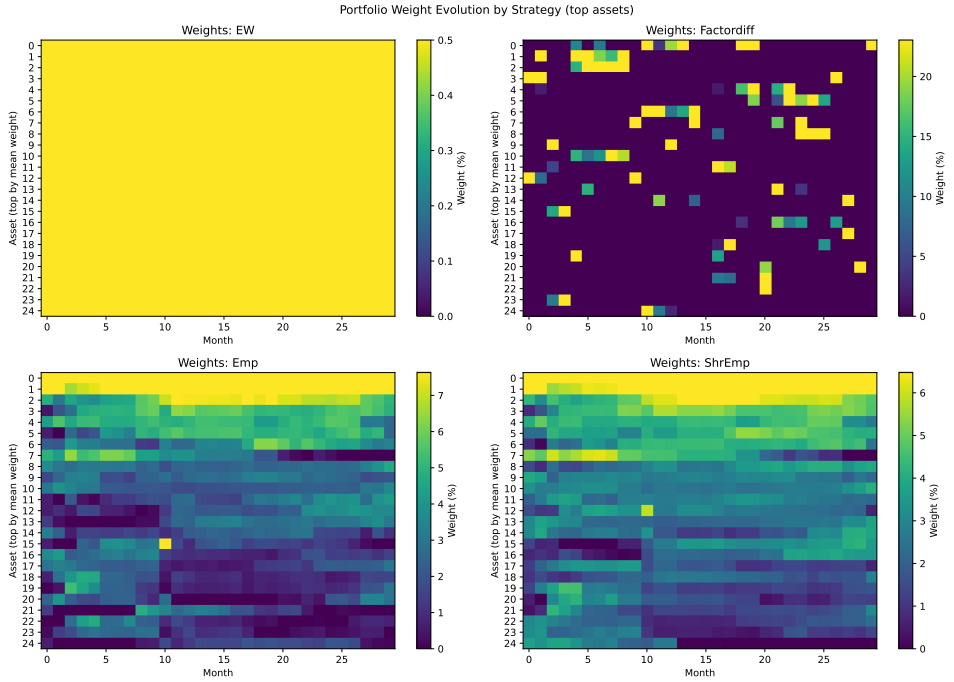


Figure 30: Weights heatmap ($k = 350$)

B.3 Implicit Factor Modeling

Figure 31 illustrates the decomposition proposed by Chen et al. (2023), where observed trajectories are separated into a low-dimensional linear subspace capturing systematic structure and an orthogonal component representing idiosyncratic variation. The projection onto the linear subspace can be

interpreted as the evolution of factors, while the orthogonal space captures residual noise. Rather than explicitly specifying factors, score estimation learns this decomposition implicitly by modeling gradients of the data density along both directions. This perspective motivates implicit factor modeling, where diffusion models recover low-dimensional structure directly through score learning without requiring predefined factor exposures. Future studies should evaluate whether this approach can match the results following the factor dimension ablation in this work.

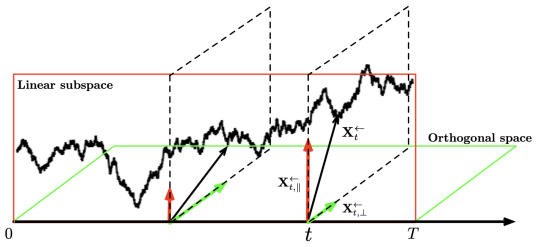


Figure 31: Image from Chen et al. (2023)