Quant Awards

The Quantitative Finance Competition

2024 winners

- 1st prize: Enhancing Portfolios Performance with Ensemble Machine Learning Models and Dynamic Thresholding – Victor Anciaux – UC Louvain (Louvain-la-Neuve, Belgium)

- 2nd prize: Dynamic Clustering in Multi-Factor Copulas with Hidden Markov Models – Anouk Rensen – Erasmus Universiteit Rotterdam (Rotterdam, Netherlands)

- 3rd prize: Multiscale Inefficiency Index – Alexandre Remiat – Université Paris Dauphine PSL (Paris, France)

Background information

More than a century after the seminal work of Louis Bachelier, the quantitative approach to financial markets has become omnipresent.

Today, many investment firms specialize in researching, developing, and implementing systematic trading strategies, increasingly incorporating data science and AI, while other active managers are adding quantitative approaches to their offerings.
Individual clients, too, can now delegate portfolio management to robo-advisors.

These examples illustrate a shifting landscape, where we believe quantitative portfolio management will continue to grow in importance due to the discipline of scientific methods and the full automation of the investment process.

The QuantAwards competition is open to university students and interns in quantitative finance regardless of specialty. Applications are free. No CFA candidacy or membership is required.

This competition offers students a unique opportunity to showcase their creativity and their understanding of this highly timely subject and to benefit from enhanced visibility in a competitive landscape.

The QuantAwards are organized and run by volunteers from eight CFA Societies, the local associations of CFA charterholders, affiliated with CFA Institute :
CFA Society Belgium, CFA Society Italy, CFA Society France, CFA Society Netherlands, CFA Society Ireland, CFA Society Norway, CFA Society Istanbul and CFA Society Spain

CFA Quant Awards 2025

Enhancing Portfolios Performance with Ensemble Machine Learning Models and Dynamic Thresholding

Abstract

Forecast combinations have proven valuable in diverse predictive contexts. Yet, their application to stock return classification remains underexplored. This paper extends the model combination framework of Roccazzella et al. (2022) from regression to classification tasks, with a focus on its integration into investment decision-making. Using NASDAQ-100 constituents from 2007 to 2024 in a rolling window setting, multiple machine learning classifiers are trained to predict weekly stock outperformance relative to the index. The continuous ensemble outputs are converted into trading signals through an optimized classification threshold, whose calibration explicitly accounts for asymmetric error costs. The methodology is evaluated across 544 portfolio configurations, spanning 4 distinct investment strategies, 4 threshold strategies, 21 ensemble methods and accounting for realistic transaction costs. All of them are evaluated through a comprehensive and robust statistical framework measuring both predictive and financial performance. Empirical results indicate that the proposed combination approach improves classification accuracy and portfolio efficiency. In particular, we find that a filtered mean–variance allocation with Ledoit–Wolf shrinkage and a precision-optimized threshold achieved a Sharpe ratio of **1.11** and an APY of **19.29**%, compared with **0.74** and **14.72**% for the benchmark. These findings highlight the practical relevance of machine learning ensemble methods with adaptive thresholding in the context of investment strategies.

1 Introduction

Predicting financial markets is known to be a major challenge. For instance, among large-cap funds, 75.25% underperformed the S&P 500 over five-year horizons, rising to 84.34% and 89.5% over ten- and fifteen-year periods (S&P Global, 2024). These results highlight the challenges of financial forecasting and the need for models capable of managing the vast and dynamic data of modern financial markets. Traditional statistical methods often fail to capture this complexity. With the rapid growth in data availability, machine learning techniques have become powerful tools to improve forecasts accuracy (Graf et al., 2020). Financial markets provide an especially rich testing ground; they produce enormous streams of heterogeneous data whose complexity and non-linearity are well suited for advanced machine learning models. However, performance varies considerably across models: Each captures certain patterns while overlooking others, and single models often lack robustness when used in isolation (Htun et al., 2024).

Ensemble learning offers a way to overcome these limitations by combining different models, capturing diverse signals in financial data and improving stability and accuracy, ultimately leading to a form of ”model consensus” (Atiya, 2020). Intuitively, this reduces model risk through diversification, similar to modern portfolio theory, offering a familiar rationale for model combination (Markowitz, 1952). Building on Roccazzella et al. (2022), this paper extends regression-based ensembles to a classification setting, where aggregated outputs are continuous probabilities of outperformance. This raises the key question of how to turn such continuous forecasts into binary investment decisions, a crucial step for portfolio construction.

Our purpose is to use ensemble machine learning models as a stock selection mechanism that identifies assets with a high probability of near-term outperformance, and to compare portfolio strategies built on this reduced set to those based on the full index. Specifically, we consider the NASDAQ-100 as investment universe and proceed as follows. Multiple classifiers featuring a large set of 34 technical indicators are fitted to each NASDAQ-100 constituent i. Each classifier j yields a binary output $\hat{y}_i^{(j)} \in \{0, 1\}$ indicating whether asset i is expected to outperform or not its index. Second, we aggregate the m binary signals through a weighted average scheme, where the vectors of combining weights $\mathbf{w}$ are optimized as described above. Third, we convert this average into a binary variable using a well-chosen cut-off threshold, and use it as a membership indicator for the asset. Fourth, having done this for each asset of the index, we apply well-known investment strategies (such as equally-weighted or mean-variance) that will assign an investment weight (γ_i) to each selected asset.

This design enables a comprehensive evaluation of ensemble-based stock selection, comparing 544 portfolio configurations to a benchmark index in realistic trading conditions. The paper is organized as follows. First we introduce forecast combination methods in Section 2. Section 3 presents the thresholding approach. Section 4 describes the portfolio optimization framework. Section 5 details the methodology. Section 6 reports the empirical results, and Section 7 concludes.

2 Ensemble Modeling Approaches

Forecast combination methods can be classified according to how they group individual forecasts, assign weights, and whether learning takes place. In this thesis, the combination problem follows the optimization framework of Roccazzella et al. (2022), treating forecasts analogously to risky assets in a portfolio. While originally designed for regression tasks, we extend this to classification, where aggregated probabilities are binarized via a threshold τ.

To formalize our idea, let $\widehat{\mathbf{Y}} = [\widehat{\mathbf{y}}^{(1)}, \ldots, \widehat{\mathbf{y}}^{(m)}] \in \mathbb{R}^{N \times m}$ be the forecast matrix, $\mathbf{w} \in \mathbb{R}^m$ the non-negative weight vector with $\sum_{j=1}^m w_j = 1; w_j \geq 0 \iff w_j \in \mathcal{W}^+$, and $\widehat{\mathbf{y}}^{\text{agg}} = \widehat{\mathbf{Y}}\mathbf{w}$ the combined forecast. $\mathcal{M} = \{1, \ldots, m\}$ is the set of all individual models and $\mathcal{W}^+$ is the non-negative weights space. The goal is to find $\mathbf{w}$ minimizing forecast error over a set of observations. The strategies described below differ in their weights selection.

The forecast combination methods applied in this paper can be grouped into several categories. A first group consists of averaging-based methods. The Simple Average (SA) assigns equal weights $w_j = 1/m$ to all models, providing a robust benchmark against overfitting (Genre et al., 2013; Smith & Wallis, 2009). A weighted variant, the Inverse-Loss (IL), assigns weights proportional to $1/\bar{e}_j$, where $\bar{e}_j$ is the mean classification error (0–1 loss) over a validation set. This approach favors models with lower misclassification rates and maintains consistency with the covariance-based methods described later. A second group includes trimming and screening-based methods, which remove poor or redundant models before averaging. In Fixed Trimming (T-SA-p), the p worst models according to mean error are excluded. The Cross-Validated Trimming (T-SA-cv) variant selects p through validation, while the Model Confidence Set (T-MCS-α) retains only models statistically different from the best at a given confidence level α (Hansen et al., 2011), using a bootstrap-based simplification (Roccazzella et al., 2022). Another important category is Constrained Optimization (CO), which, comparable to minimum-variance portfolio allocation, minimizes $\mathbf{w}^\top \boldsymbol{\Sigma} \mathbf{w}$ under simplex constraints, where $\boldsymbol{\Sigma}$ is the empirical covariance matrix of forecast errors. This framework is extended in Penalized Constrained Optimization (COP), which shrinks $\mathbf{w}$ toward a reference weight vector $\mathbf{w}^{\text{ref}}$ (either equal or inverse-loss weights) via a penalty term $d(\mathbf{w}, \mathbf{w^0})$, implemented as L1, L2, elastic net, or cross-entropy penalties (Roccazzella et al., 2022). This improves robustness, particularly in small-sample settings. Finally, Covariance Shrinkage Optimization (COS) stabilizes $\boldsymbol{\Sigma}$ by shrinking it toward a structured target $\boldsymbol{\Sigma}^0$. In the COS-E variant, $\boldsymbol{\Sigma}^0 = \overline{\sigma}^2 \mathbf{I}$ assumes equal, uncorrelated errors. In COS-IL, $\boldsymbol{\Sigma}^0$ is a diagonal matrix with entries σ_j^2 given by the individual model error variances, approximating inverse-loss weighting. Final weights are obtained by solving $min_{\mathbf{w} \in \mathcal{W}^+} \mathbf{w}^\top \boldsymbol{\Sigma}_\lambda \mathbf{w}$, with the shrinkage intensity λ tuned via cross-validation. Table 1 summarizes the 17 unique combinations methods.

Combination methods	**Acronym**	**Weights computation**
Standard methods		
Simple average forecast	SA	$\overline{\mathbf{w}}^E$
Loss-based weighted average forecast	IL	$\overline{\mathbf{w}}^{IL}$
Constrained optimization	CO	$\psi(\boldsymbol{\Sigma})$
Trimmed averages		
Trimmed simple average	T-SA-p	$1/\text{card}(\mathcal{M}_p)$ if $i \in \mathcal{M}_p$, 0 otherwise
Trimmed simple average via cross-validation	T-SA-CV	$1/\text{card}(\mathcal{M}_p)$ if $i \in \mathcal{M}_p$, 0 otherwise
Trimmed-MCS simple average	T-MCS-α	$1/\text{card}(\mathcal{M}_\alpha)$ if $i \in \mathcal{M}_\alpha$, 0 otherwise
CO with shrinkage to $\overline{\mathbf{w}}^E$		
COP with L1 penalty	COP_L1^E	$\arg\min_{\mathbf{w}} \mathbf{w}^\top \Sigma \mathbf{w} + \delta \lVert \mathbf{w} - \overline{\mathbf{w}}^E \rVert_1$
COP with L2 penalty	COP_L2^E	$\arg\min_{\mathbf{w}} \mathbf{w}^\top \Sigma \mathbf{w} + \delta \lVert \mathbf{w} - \overline{\mathbf{w}}^E \rVert_2^2$
COP with EN penalty	COP_EN^E	$\arg\min_{\mathbf{w}} \mathbf{w}^\top \Sigma \mathbf{w} + \delta \left[\alpha \lVert \mathbf{w} - \overline{\mathbf{w}}^E \rVert_1 + (1-\alpha) \lVert \mathbf{w} - \overline{\mathbf{w}}^E \rVert_2^2 \right]$
COP with EN penalty + Shrinkage to 0	COP_EN	$\arg\min_{\mathbf{w}} \mathbf{w}^\top \Sigma \mathbf{w} + \delta \left[\alpha \lVert \mathbf{w} \rVert_1 + (1-\alpha) \lVert \mathbf{w} \rVert_2^2 \right]$
COP with cross-entropy	COP_CE^E	$\arg\min_{\mathbf{w}} \mathbf{w}^\top \Sigma \mathbf{w} + \delta \sum_{i=1}^m w_i \ln \left(\frac{w_i}{\overline{\mathbf{w}}_i^E} \right)$
CO shrinkage with reference $\overline{\mathbf{w}}^E$	COS^E	$\psi(\boldsymbol{\Sigma}^E_{\widehat{\lambda^*}})$
CO with shrinkage to $\overline{\mathbf{w}}^{IL}$		
COP with L1 penalty	COP_L1^{IL}	$\arg\min_{\mathbf{w}} \mathbf{w}^\top \Sigma \mathbf{w} + \delta \lVert \mathbf{w} - \overline{\mathbf{w}}^{IL} \rVert_1$
COP with L2 penalty	COP_L2^{IL}	$\arg\min_{\mathbf{w}} \mathbf{w}^\top \Sigma \mathbf{w} + \delta \lVert \mathbf{w} - \overline{\mathbf{w}}^{IL} \rVert_2^2$
COP with EN penalty	COP_EN^{IL}	$\arg\min_{\mathbf{w}} \mathbf{w}^\top \Sigma \mathbf{w} + \delta \left[\alpha \lVert \mathbf{w} - \overline{\mathbf{w}}^{IL} \rVert_1 + (1-\alpha) \lVert \mathbf{w} - \overline{\mathbf{w}}^{IL} \rVert_2^2 \right]$
COP with cross-entropy	COP_CE^{IL}	$\arg\min_{\mathbf{w}} \mathbf{w}^\top \Sigma \mathbf{w} + \delta \sum_{i=1}^m w_i \ln \left(\frac{w_i}{\overline{\mathbf{w}}_i^{IL}} \right)$
CO shrinkage with reference $\overline{\mathbf{w}}^{IL}$	COS^{IL}	$\psi(\boldsymbol{\Sigma}^E_{\widehat{\lambda^*}})$

Table 1: List of the considered forecast combination methods. *Minimization is performed over $\mathcal{W}^+$. The notation card$(\cdot)$ denotes the cardinality of a set.* $\boldsymbol{\Sigma}$ denotes the covariance matrix of the errors, computed trough the 10% combination test set. For the elastic net divergence measure, we set $\alpha = 0.8$ based on Roccazzella et al. (2022). We estimate the penalties δ via 5-fold cross-validation. $p \in \{1, 2, 3, 4\}$ and $\alpha \in \{0.2, 0.5\}$

.

3 From Regression-Based Ensembles to Classifier Ensembles: A Thresholding Approach

Most of the forecast combination literature, including the framework of Roccazzella et al. (2022), has been developed in a **regression** setting, where models output continuous point forecasts. In such contexts, aggregation produces another continuous value that can be directly compared to the target, with no post-processing step.

In this paper, the framework is extended to a classification setting, where each model predicts the outperformance of future stock returns relative to the index. Individual predictions $\hat{y}_i^{(j)} \in \{0, 1\}$ are aggregated as: $\hat{y}_i^{\text{agg}} = \sum_{j=1}^m w_j \, \hat{y}_i^{(j)}$ This yields a continuous score in $[0, 1]$. To produce final binary trading signals, a classification threshold $\tau \in [0, 1]$ is applied:

$$\widehat{y_i}^{\text{class}} = \begin{cases} 1 & \text{if } \widehat{y_i}^{\text{agg}} \geq \tau \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

A fixed $\tau = 0.5$ is common in classification, assuming balanced classes and calibrated probabilities. However, these assumptions rarely hold in financial prediction, where data is often imbalanced and the cost of false positives and false negatives might be asymmetric. An unsuitable threshold can reduce predictive power and harm portfolio performance.

To address this, we treat threshold selection as a post-combination optimization problem. Four strategies are implemented:

- Naïve Threshold (`naive`): Default $\tau = 0.5$, as used in (Simonis, 2021), appropriate only under balanced and symmetric conditions.
- Direct Optimization (`OPTI`): Selects $\tau^* = \arg\max_\tau \mathcal{C}_\mathcal{M}(\tau)$, where $\mathcal{C}_\mathcal{M}$ is a chosen metric (e.g., F1-score, accuracy) computed on a validation set.
- Cross-Validated Thresholding (`CV`): Applies TimeSeriesSplit cross-validation to choose τ that maximizes average metric performance across folds, reducing overfitting in small validation sets.
- Quantile-Based Thresholding (`QTL`): Sets τ based on the empirical distribution of aggregated scores (e.g., 75th percentile), useful when no ground truth is available for tuning.

By explicitly optimizing τ, classification ensembles can be better aligned with financial decision-making goals, adapting to imbalanced datasets and asymmetric risk preferences.

4 Portfolio Optimization

We adopt a classification-based approach: Each year, machine learning models predict stocks likely to outperform the index the following week. Capital is then allocated among selected assets using four strategies: equal weight (`EW`), mean variance(`MV`), shrinkage (`SHRINKAGE`) and hybrid shrinkage–filtering (`FILTERING`).

The `EW` portfolio assigns $\gamma_i = 1/N$ to each asset (where N is the number of asset predicted as out-performer), a robust benchmark often beating optimized portfolios (DeMiguel et al., 2009). In `MV` (Markowitz, 1952), variance $\gamma^\top \hat{\mathbf{\Sigma}} \gamma$ is minimized using $\hat{\mu} = \frac{1}{T}\sum_{t=1}^{T} r_t$ and $\hat{\mathbf{\Sigma}} = \frac{1}{T-1}\sum_{t=1}^{T}(\mathbf{r_t} - \hat{\mu})(\mathbf{r_t} - \hat{\mu})^\top$, the empirical covariance of forecast returns $\mathbf{r_t}$, where T is the number of periods, applied only to predicted outperformers. As these estimates are noisy and unstable when N is large (Jagannathan & Ma, 2003; Michaud, 1989), we apply two regularizations based on the empirical covariance of the returns:

(i) Shrinkage (Ledoit & Wolf, 2004) blends $\hat{\Sigma}$ with a structured target $\mathbf{\Theta}$: $\widehat{\mathbf{\Sigma}}^{LW} = \delta\mathbf{\Theta} + (1 - \delta)\widehat{\mathbf{\Sigma}}$, with δ minimizing MSE (Ledoit & Wolf, 2003).

(ii) Filtering (Bun et al., 2017; Pafka et al., 2004) uses RMT to remove noise eigenvalues in $[\lambda_-, \lambda_+]$ from $\widehat{\mathbf{C}} = \mathbf{D}^{-1}\hat{\mathbf{\Sigma}}\mathbf{D}^{-1}$, retaining $\widehat{\mathbf{C}}^{(s)}$ and reconstructing $\widehat{\mathbf{\Sigma}}^{filtered} = \mathbf{D}\,\widehat{\mathbf{C}}^{(s)}\,\mathbf{D}$.

The Hybrid method shrinks $\widehat{\mathbf{\Sigma}}^{filtered}$ toward $\widehat{\mathbf{\Sigma}}^{LW}$: $\widehat{\mathbf{\Sigma}}^{final} = (1 - \delta)\widehat{\mathbf{\Sigma}}^{filtered} + \delta\widehat{\mathbf{\Sigma}}^{LW}$, with $\delta = 0.3$. This choice keeps the filtered covariance predominant, while allowing the Ledoit–Wolf shrinkage to absorb residual noise, thus combining spectral noise reduction and regularization for stable covariance estimation. Since we do not forecast returns, the empirical covariance of the returns are estimated with 5-years past data.

5 Methodology

We focus on the **NASDAQ-100** index from January 2007 to December 2024 using the actual historical composition year by year to avoid survivorship bias, leading to $T = 936$ periods. Constituents are fixed at the start of each year, and only stocks with sufficient historical data are included. All market and composition data are sourced from Bloomberg L.P. (2024).

Technical analysis features are based solely on price and volume, computed from weekly data aggregated from daily closes and volumes. In total, 34 technical indicators are used (Table 4), covering moving averages, momentum measures, volatility proxies, regression slopes, skewness and volume-based indicators, all previously identified in seminal papers (Kakushadze, 2016; Wolff & Echterling, 2022).

A walk-forward procedure with rolling five-year windows is used: 70% training, 10% validation (ensemble tuning and post-computation of threshold τ), 20% testing (Figure 5). The set of $m = 10$ individual models comprises Ridge, Lasso, ElasticNet, DecisionTree, RandomForest, GradientBoosting, XGBoost, SVM, KNN and MLP. Hyperparameters are tuned via randomized grid search over 10 parameter sets with 2-fold time-series CV on the training portion. Each year, the window advances by one year, producing 52 evaluations per stock each year from 2007–2024. All preprocessing steps, standardization and a principal component analysis retaining 95% of variance, are applied strictly within each rolling training window to avoid look-ahead bias. The resulting transformed dataset, denoted $\tilde{\mathbf{X}}$, is then used as model input. Computations leverage GPU acceleration via CuML.

Concretely, our approach proceeds as following. First, individual ML classifiers are trained on the technical features of each asset to predict whether its weekly return will outperform the benchmark, yielding binary buy/no-buy signals. Second, these signals are aggregated through various ensemble and thresholding strategies to produce a consolidated selection. Third, assets flagged with a positive signal are allocated using alternative portfolio construction rules (equal-weighting, mean–variance with Ledoit–Wolf shrinkage, filtering, etc.), while accounting for transaction costs at each rebalancing (Interactive Brokers' commission schedule Interactive Brokers (2024): \$0.005 per share with minimum \$1 and maximum 1% of trade value). Since we do not forecast returns directly, the covariance matrix $\widehat{\Sigma}$ is estimated from the past five years of realized returns. Performance is then benchmarked against a NASDAQ-100 ETF (Invesco, 2024), with the 3-month Euribor as risk-free rate. Robustness is assessed by varying sample periods, transaction costs, and hyperparameters criteria. Full results of those robustness tests are reported in Appendix.

6 Empirical Results

We now assess the practical value of the predictive models by evaluating their performance in portfolio construction. The key question is whether better classification can translate into superior financial gains. We test four allocation strategies, Equal Weight (`EW`), Mean–Variance Optimization (`MV`), shrinkage-based MV, and an hybrid method combining filtering and shrinkage (`FILTERING`), applied to both individual and ensemble forecasts, under multiple thresholding rules. We compare all results with a simple passive investment in the NASDAQ-100 index. In total, the analysis covers 21 ensemble methods, 10 individual classifiers, 6 thresholding methods, and 4 allocation schemes, resulting in 544 distinct portfolio strategies. We first analyze the classification performance of individual models and model combinations for the naïve thresholding scheme. Each result is statistically tested against a random baseline, defined as the performance of a model predicting 1 or 0 with equal probability (50%). Given the dataset's class imbalance (48.5% of positive cases), the baseline performance naturally reflects this initial distribution. Next, we investigate the impact of the thresholding strategy on the various classification metrics. Finally, we discuss the financial performance of the associated setups in light of our Nasdaq Benchmark.

Table 2 reports the performance metrics of our combination schemes when adopting the naive classification threshold ($\tau = 0.5$). One can see that many ensemble methods outperform the simple average (`SA`) in accuracy, though `SA` remains among the top in F1-score, consistent with literature on the robustness of equal weights (DeMiguel et al., 2009; Goyal & Welch, 2008). Rankings vary by metric: `T-SA-4`, for example, is the only method above statistical significance thresholds for all four metrics, reflecting balance and robustness.

Model	Accuracy	Precision	Recall	F1	McNemar
T-SA-3	**50.97%*****	**49.17%*****	44.67%	46.82%	***
IL	50.82%***	49.01%***	43.78%	46.25%	***
T-SA-1	50.88%***	48.99%**	43.51%	46.09%	**
T-SA-CV	50.78%***	49.01%***	45.69%	47.29%	***
COP_L1^{E}	50.77%***	48.99%**	45.69%	47.29%	***
COS^{E}	50.76%***	48.97%**	44.90%	46.85%	**
COP_L1^{IL}	50.71%***	48.89%*	43.87%	46.24%	***
COP_EN	50.71%***	48.92%**	45.22%	47.00%	***
COS^{IL}	50.70%***	48.86% *	43.67%	46.12%	***
COP_L2^{E}	50.69%***	48.89%*	44.54%	46.61%	***
COP_L2^{IL}	50.69%***	48.88%*	44.53%	46.61%	**
COP_CE^{IL}	50.69%***	48.86%*	43.89%	46.25%	**
COP_EN^{IL}	50.67%***	48.85%*	44.06%	46.33%	**
CO	50.67%***	48.88%*	45.63%	47.20%	**
COP_CE^{E}	50.66%***	48.83%*	43.90%	46.23%	**
COP_EN^{E}	50.64%***	48.81%*	44.16%	46.37%	/
SA	50.64%***	48.92%**	48.77%	48.84%	**
T-SA-2	50.63%***	48.93%**	49.50%	49.21%	***
T-SA-4	50.61%***	48.97%**	**52.09%*****	**50.48%****	*
T-MCS-50	50.47%**	48.79%*	50.02%	49.39%	**
T-MCS-20	50.35%*	48.51%	44.61%	46.48%	/

Model	Accuracy	Precision	Recall	F1	McNemar
RandomForest	**50.89%*****	**49.13%*****	45.64%	47.32%	***
Elasticnet	50.89%***	49.10%***	43.76%	46.27%	***
Lasso	50.89%***	49.09%***	43.77%	46.28%	***
Ridge	50.89%***	49.09%***	43.76%	46.27%	***
MLP	50.58%***	48.84%*	47.91%	**48.38%**	***
GradientBoosting	50.52%**	48.78%*	47.41%	48.08%	**
SVM	50.42%**	48.58%	44.32%	46.35%	*
KNN	50.34%*	48.57%	46.84%	47.69%	-
DecisionTree	50.19%	48.45%	48.14%	48.29%	-
XGBoost	50.18%	48.44%	48.13%	48.29%	-

Table 2: Performance metrics and statistical tests for each individual and ensemble model using a naive threshold strategy, *based on aggregated out-of-sample predictions. Statistical significance stars ('***', '**', '*') indicate results that are statistically significant at the 0.1%, 1%, and 5% levels respectively, based on one-sided binomial tests against a random baseline classifier (accuracy = 50%, precision = 48.5%, recall = 50%, directly due to proportion of ones and zeros in the initial dataset (48.5% of 1)). The final column reports McNemar test significance, evaluating whether each model's prediction pattern significantly differs from random guessing.*

Interestingly, looking beyond average scores reveals an important risk associated with relying only on individual models. Despite strong average performance, individual models display pronounced year-to-year variability. The results shown in Figure 1 reveal a marked dispersion in annual performance: A model that excels in one year may underperform in the next, likely reflecting shifts in market dynamics or data structure. This illustrates what is commonly referred to as model risk: The danger that relying on a single model may lead to poor outcomes if that model happens not to be the appropriate one at a given time. Such variability highlights the value of combining forecasts, which mitigates the dependence on any single specification.

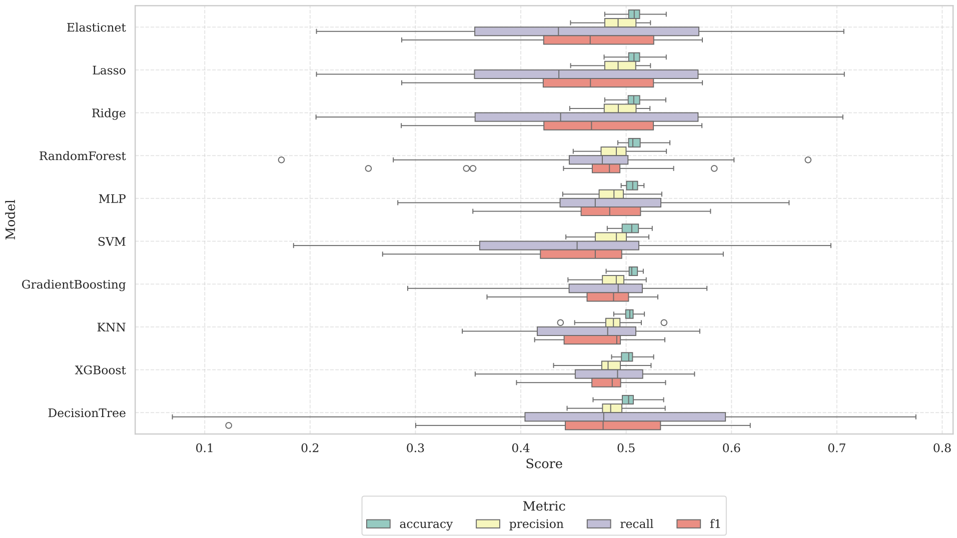

Figure 1: Boxplot of classification metrics for each individual model. *The models are evaluated over all out-of-sample periods using a walk-forward validation scheme. This representation highlights both the central tendency and the dispersion of predictive performance across models and time. Indeed, more information about the dispersion of the metrics in front of the years are available in Figure 8.*

Acknowledging this dispersion, Figure 2 illustrates how the average weight assigned to each individual model evolves over time across ensemble strategies. While the relative importance of individual models shifts drastically from year to year, the overall distribution of weights remains relatively balanced and close, but different, to the equal–weight benchmark. This suggests that the ensemble framework not only adjusts dynamically to evolving patterns, but also preserves a level of stability that prevents over–reliance, hence over confidence, on any single classifier.

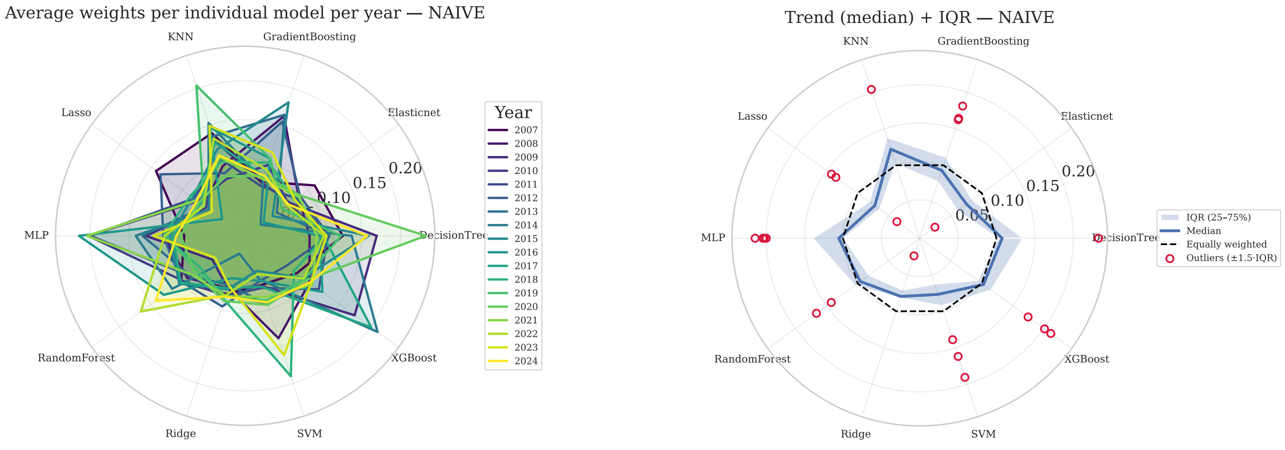

Figure 2: Average weights assigned by the ensemble models to each individual model. *(left) Year-by-year profiles: Each axis is a base model and each colored polygon is one year (2007–2024); larger radii indicate higher average weights. (right) Robust summary: Median profile with interquartile band (25–75%) and, when shown, outliers beyond* $Q_1 - 1.5\,\mathrm{IQR}$ *and* $Q_3 + 1.5\,\mathrm{IQR}$.

Before moving to the portfolio construction stage and the analysis of financial performance, we examine the role of threshold selection. Rather than limiting the evaluation to the naive $\tau = 0.5$ rule, we explore more informed strategies to understand how this choice can influence classification outcomes and, ultimately, the quality of the signals feeding into our portfolio models. Our findings confirm that threshold selection is far from a trivial detail, it's a key factor in improving classification results for ensembles. While the naive threshold of $\tau = 0.5$ offers a convenient baseline, our results demonstrate that more informed strategies can lead to meaningful performance gains. Figure 3 shows that `OPT`, `QTL`, and `CV` generally outperform the naive strategy in terms of accuracy and precision. However, this improvement often comes at the cost of significantly reduced recall, and consequently, a lower F1 score.

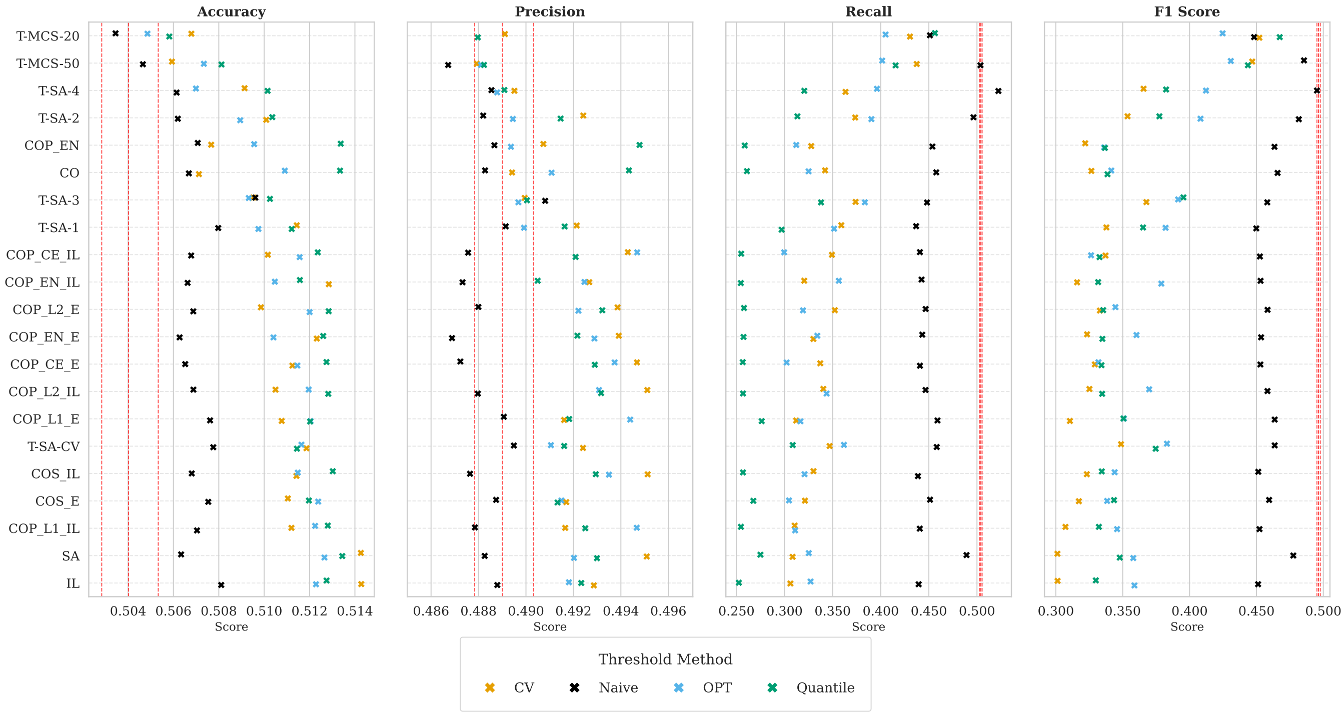

Figure 3: Performance of thresholding strategies across ensemble models and evaluation metrics. *Each cross indicates the score of a method under a specific threshold. Dashed vertical lines show significance levels (5%, 1%, 0.1%) from a one-sided binomial test under the null hypothesis of a random classifier (accuracy = recall = 50%, precision = 48.5%).* `CV` *and* `OPTI` *denote accuracy-optimized versions.*

In terms of portfolios performances, Table 3 lists the 15 highest Sharpe Ratio portfolios. All are ensemble-based and above all, we note that no individual models, under any portfolio allocation strategy, manage to outperform the benchmark. Our Hybrid-filtering (`FILTERING`) strategy dominates the top rankings, reflecting its ability to stabilize covariance estimation and enhance risk-adjusted returns. The best-performing portfolio is `T-SA-1` with precision-optimized threshold ($\texttt{OPT}_{\texttt{PREC}}$), achieving SR = 1.11 and APY = 19.29%. The simple average (`SA`) ranks third, confirming again that literature findings that equal-weight ensemble can rival complex weighting schemes. $\texttt{COP\_EN}^{E}$ completes the top three. Threshold choice is decisive: Among the top eight portfolios, all use a strategy where the threshold is optimizing precision ($\texttt{OPT}_{\texttt{PREC}}$). Figure 4 evaluates thresholding schemes in portfolio performance. While naive $\tau = 0.5$ yields the highest proportion of SR outperformance cases, precision-optimized thresholds ($\texttt{OPT}_{\texttt{PREC}}$) achieve the highest absolute APY outperformance cases. Recall-optimized thresholds ($\texttt{OPT}_{\texttt{REC}}$) lead in win rate, but at the cost of lower precision and returns.

All of our findings reinforce the financial relevance of ensemble methods as a ”model consensus”, which aggregate complementary signals and mitigate the weaknesses of single predictors. In addition to being a decisive driver of ensemble performance, thresholding strategy is also a key factor in enhancing financial performance results.

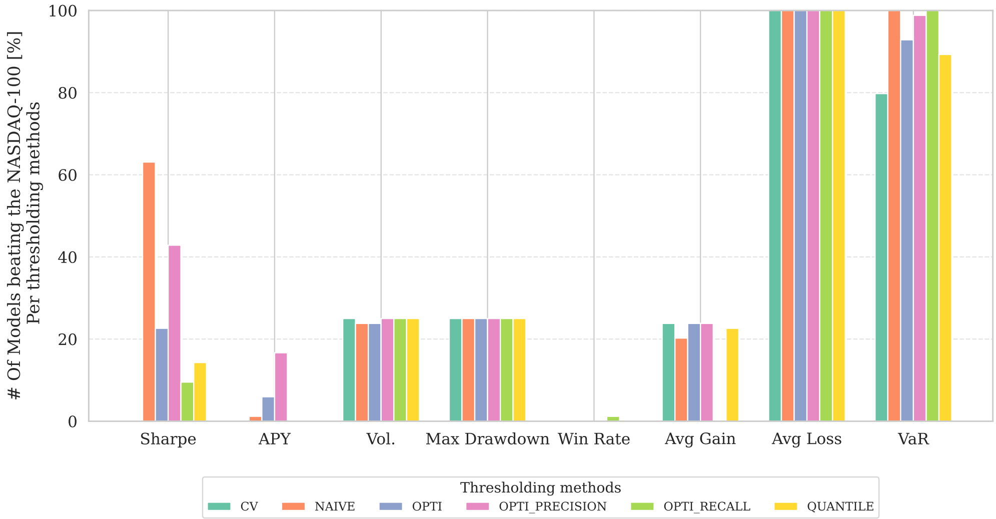

Figure 4: Percentage of portfolios beating the NASDAQ-100 by thresholding method.

Portfolio	Ensemble Method	Threshold	SR	APY [%]	Vol. [%]	DD [%]	VaR [%]	CVaR	WR [%]
FILT.	**T-SA-1**	**OPTI$_{PREC}$**	**1.11**	**19.29**	**16.26**	**-29.30**	**-3.19**	**-5.02**	**59.59**
FILT.	COP_EN^{E}	OPTI$_{PREC}$	1.10	19.07	16.24	-29.68	-3.17	-4.98	58.96
FILT.	SA	OPTI$_{PREC}$	1.08	18.34	15.86	-29.64	-3.07	-4.71	57.14
FILT.	IL	OPTI$_{PREC}$	1.08	18.53	16.10	-30.19	-3.16	-4.81	57.36
FILT.	COP_EN^{IL}	OPTI$_{PREC}$	1.00	16.78	15.84	-30.32	-3.24	-4.87	57.78
FILT.	COP_L1^{IL}	OPTI$_{PREC}$	1.00	17.25	16.34	-31.17	-3.42	-4.96	57.68
FILT.	T-SA-3	OPTI$_{PREC}$	0.99	16.75	16.00	-38.91	-3.18	-5.02	58.21
FILT.	T-SA-CV	OPTI$_{PREC}$	0.97	16.80	16.47	-38.90	-3.36	-5.18	58.53
FILT.	IL	OPTI$_{ACC}$	0.97	17.08	16.81	-35.37	-3.28	-5.30	58.21
MV	T-SA-1	OPTI$_{PREC}$	0.95	15.74	15.59	-36.19	-3.09	-5.06	59.17
FILT.	T-SA-3	NAIVE	0.95	15.60	15.54	-38.12	-3.06	-4.95	58.42
FILT.	SA	OPTI$_{ACC}$	0.95	16.73	16.79	-34.69	-3.35	-5.29	57.68
FILT.	T-SA-2	NAIVE	0.93	14.72	14.99	-31.61	-3.03	-4.71	59.06
SHRINK.	T-SA-1	OPTI$_{PREC}$	0.93	15.24	15.61	-36.56	-3.26	-5.16	59.81
MV	COP_EN^{E}	OPTI$_{PREC}$	0.92	15.25	15.68	-37.43	-3.32	-5.08	59.17
INDEX NASDAQ-100			***0.74***	***14.72***	***19.7***	***-51.53***	***-4.66***	***-6.79***	***61.73***

Table 3: Top 15 of portfolios ranking by their Sharpe Ratio.

7 Conclusion

This papers examined how a regression-based ensemble learning framework could be adapted to classification tasks aimed at predicting stock returns and building portfolios that outperform the market. Using NASDAQ-100 data and only technical indicators, we implemented individual classifiers and ensemble methods in a walk-forward setting, introducing post-combination threshold selection as a central methodological element. The results show that although some individual models did well in classification performance, **none** managed to deliver portfolios that beat the NASDAQ-100 due to their high perfromance dispersion across the years. Precision-optimized thresholds yielded the best financial results, with the top strategy (a mean-variance allocation with filtering and Ledoit–Wolf shrinkage and a one-trimmed ensemble) reaching a Sharpe Ratio of **1.11** and APY of **19.29%**, compared to **0.74** and **14.72%** for the benchmark. The choice of threshold turned out to be a key factor, affecting both prediction quality and portfolio performance. Fewer than 20% of all portfolios outperformed the benchmark in APY, underscoring market noise and difficulty of signal extraction. Different thresholds aligned with different objectives: $\texttt{OPTI}_{\texttt{PREC}}$ maximized APY, while $\tau = 0.5$ yielded more Sharpe Ratio outperformance. Risk metrics like maximum drawdown and volatility are less affected by thresholding than by the choice of portfolio allocation method. In sum, ensembles generally surpassed individual models by providing greater robustness and stability over time.

While our results are promising, several limitations point to areas for future research. First, our models rely solely on technical indicators and could benefit from the integration of fundamental or alternative data, as well as more diversified architectures (e.g., LSTMs). Second, backtests simplify real-world constraints by assuming fractional shares, no taxation on profits, and perfect execution at closing prices, which are the standard "academic" setup. Third, we adopt a classification framework leading to aggregate forecasts that need thresholds strategies. A comparison with a regression-based approach, which would predict returns directly and computing signal relative to the benchmark returns then (and so bypassing the need of a threshold strategy) remains an important avenue for future researches. Finally, our asset pre-selection is performed independently for each stock, considering only marginal features at this stage. As shown by Vanderveken et al. (2024), simply reducing the investment universe, even by random selection, can already improve performance by mitigating the risk associated with high-dimensional covariance matrix estimation. Future work should therefore compare forecast-based selection with random sub-sampling of the universe in order to distinguish the advantages of informative signals from those due solely to the dimension reduction.

Overall, this paper highlights that, in a classification framework, ensemble learning when combined with carefully calibrated thresholds and robust portfolio design can lead to significant improvements in both statistical and financial terms. Among the various design choices, the threshold mechanism stands out as one of the most influential factors, as it directly determines how predictive signals are translated into investment actions, which ultimately impacts portfolio performance.

References

Atiya, A. F. (2020). Why does forecast combination work so well? [M4 Competition]. *International Journal of Forecasting*, *36*(1), 197–200. https://doi.org/https://doi.org/10.1016/j.ijforecast.2019.03.010

Bloomberg L.P. (2024). Bloomberg terminal historical market data [Accessed via Bloomberg Terminal].

Bun, J., Bouchaud, J.-P., & Potters, M. (2017). Cleaning large correlation matrices: Tools from random matrix theory. *Physics Reports*, *666*, 1–109.

DeMiguel, V., Garlappi, L., & Uppal, R. (2009). Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *The Review of Financial Studies*, *22*(5), 1915–1953. https://doi.org/10.1093/rfs/hhm075

Genre, V., Kenny, G., Meyler, A., & Timmermann, A. (2013). Combining expert forecasts: Can anything beat the simple average? *International Journal of Forecasting*, *29*(1), 108–121. https://doi.org/10.1016/j.ijforecast.2012.06.004

Goyal, A., & Welch, I. (2008). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, *21*(4), 1455–1508.

Graf, C., Flanagan, D., Wylie, L., & Silver, D. (2020). The Open Data Challenge: An Analysis of 124,000 Data Availability Statements and an Ironic Lesson about Data Management Plans. *Data Intelligence*, *2*(4), 554–568. https://doi.org/10.1162/dint_a_00061

Hansen, P. R., Lunde, A., & Nason, J. M. (2011). The model confidence set. *Econometrica*, *79*(2), 453–497. https://doi.org/10.3982/ECTA5771

Htun, H., Biehl, M., & Petkov, N. (2024). Forecasting relative returns for sp 500 stocks using machine learning. *Financial Innovation*, *10*. https://doi.org/10.1186/s40854-024-00644-0

Interactive Brokers. (2024, June). Commission schedule [Accessed: 2025-06]. https://www.interactivebrokers.com/en/pricing/commissions.php

Invesco. (2024). Invesco eqqq nasdaq-100 ucits etf - accumulation [Accessed June 2024].

Jagannathan, R., & Ma, T. (2003). Risk reduction in large portfolios: Why imposing the wrong constraints helps. *The Journal of Finance*, *58*(4), 1651–1684.

Kakushadze, Z. (2016). 101 formulaic alphas. https://arxiv.org/abs/1601.00991

Ledoit, O., & Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, *10*(5), 603–621.

Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, *88*(2), 365–411.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, *7*(1), 77. https://doi.org/10.2307/2975974

Michaud, R. O. (1989). The markowitz optimization enigma: Is 'optimized' optimal? *Financial Analysts Journal*, *45*(1), 31–42.

Pafka, S., Potters, M., & Kondor, I. (2004). Exponential weighting and random-matrix-theory-based filtering of financial covariance matrices for portfolio optimization. *arXiv preprint cond-mat/0402573*.

Roccazzella, F., Gambetti, P., & Vrins, F. (2022). Optimal and robust combination of forecasts via constrained optimization and shrinkage. *International Journal of Forecasting*, *38*(1), 97–116. https://doi.org/10.1016/j.ijforecast.2021.04.002

Simonis, N. (2021). *Stock selection and portfolio optimization: A machine learning approach* [Master's thesis]. Université catholique de Louvain. http://hdl.handle.net/2078.1/thesis:31397

Smith, J., & Wallis, K. F. (2009). A simple explanation of the forecast combination puzzle. *Oxford Bulletin of Economics and Statistics*, *71*(3), 331–355. https://doi.org/10.1111/j.1468-0084.2008.00541.x

S&P Global. (2024, December). SPIVA — S&P Dow Jones Indices — spglobal.com [[Accessed 02-08-2025]].

Vanderveken, R., Lassance, N., & Vrins, F. (2024). Optimal portfolio size under parameter uncertainty [Available at UCLouvain Institutional Repository]. *Working Paper*. https://dial.uclouvain.be/pr/boreal/en/object/boreal:289687

Wolff, D., & Echterling, F. (2022). Stock picking with machine learning [Available at https://ssrn.com/abstract=3607845].

Appendix

Indicator (daily data, pre-resampling)	**Period (Days)**
log(Price / SMA$_{50,100,200}$)	50, 100, 200
log(Price / EMA$_{50,100,200}$)	50, 100, 200
MACD level and signal	-
MACD crossover binary indicator	-
Relative Strength Index (RSI)	14
log(Price / Upper Bollinger Band)	20
log(Price / Lower Bollinger Band)	20
Indicator (weekly data, post-resampling)	**Period (Weeks)**
Weekly return	1
Excess return (vs index)	13 & 26
Lagged stock return	13 & 26 & 52
Price momentum	13 & 26 & 52
Beta (stock vs index)	13 & 26
Rolling return volatility	13 & 26
Return / Volatility (Sharpe proxy)	13
Return-volume correlation	13
Regression slope (linear trend of price)	13 & 26
Weekly RSI	13
Return skewness	13 & 26
On-Balance Volume (OBV)	-
Dollar trading volume	-

Table 4: Description of the technical indicators. *Indicators with a specified period are computed using a backward-looking rolling window covering the stated number of weeks. Indicators without a defined period are computed using the available data at the time of observation. All indicators based on daily data are computed prior to resampling the data to a weekly frequency. The computations are performed using the* `TA-Lib` *Python library.*

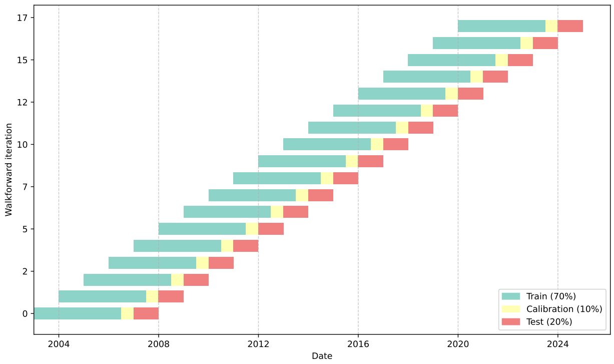

Figure 5: Walk-forward strategy for model training. *Each window is split chronologically into 70% for training base models, 10% for calibrating the threshold τ and tuning ensemble methods, and 20% for out-of-sample evaluation.*

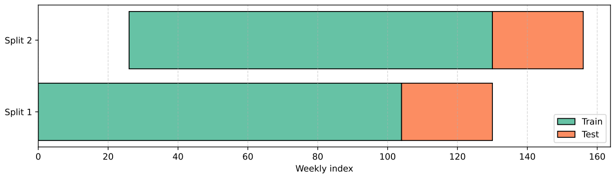

Figure 6: Detail of the hyperparameter tuning stage.*A randomized grid search over 10 sampled parameter combinations with 2-fold time-series cross-validation performed on the 70% training segment. Each training fold is composed of 2 years of observations.*

# Step	Details
1	Load historical prices and compute weekly returns.
2	Compute technical indicators (momentum, moving averages, RSI, etc.).
3	Build features matrix X and binary target y (outperform / not).
4	Split data into training, validation, and test sets (walk-forward).
5	Train individual classifiers (Logistic Regression, SVM, RF, XGBoost, NN, ...).
6	Generate outperformance forecasts for each stock and each week.
7	Combine forecasts through ensemble methods (average, trimmed, optimization, ...).
8	Apply thresholding rule to convert probabilities into binary buy/no-buy signals.

Table 5: Prediction process prior to portfolio construction.

#	Step	Details
1	Load signals	Retrieve current week's buy/no-buy signals.
2	Exit positions	Sell holdings without renewed buy signal.
3	Select tickers	Keep all flagged as buy for this week.
4	Estimate risk	Compute covariance from 4 years of weekly returns.
5	Optimize weights	• `EW`: Equal Weight • `MV`: Minimum Variance • `SHR`: Ledoit–Wolf shrinkage • `FILT`: RMT filter, shrunk 30% to Ledoit–Wolf
6	Rebalance	Weekly, including Interactive Brokers' commissions (June 2024).

Table 6: Portfolio construction process.

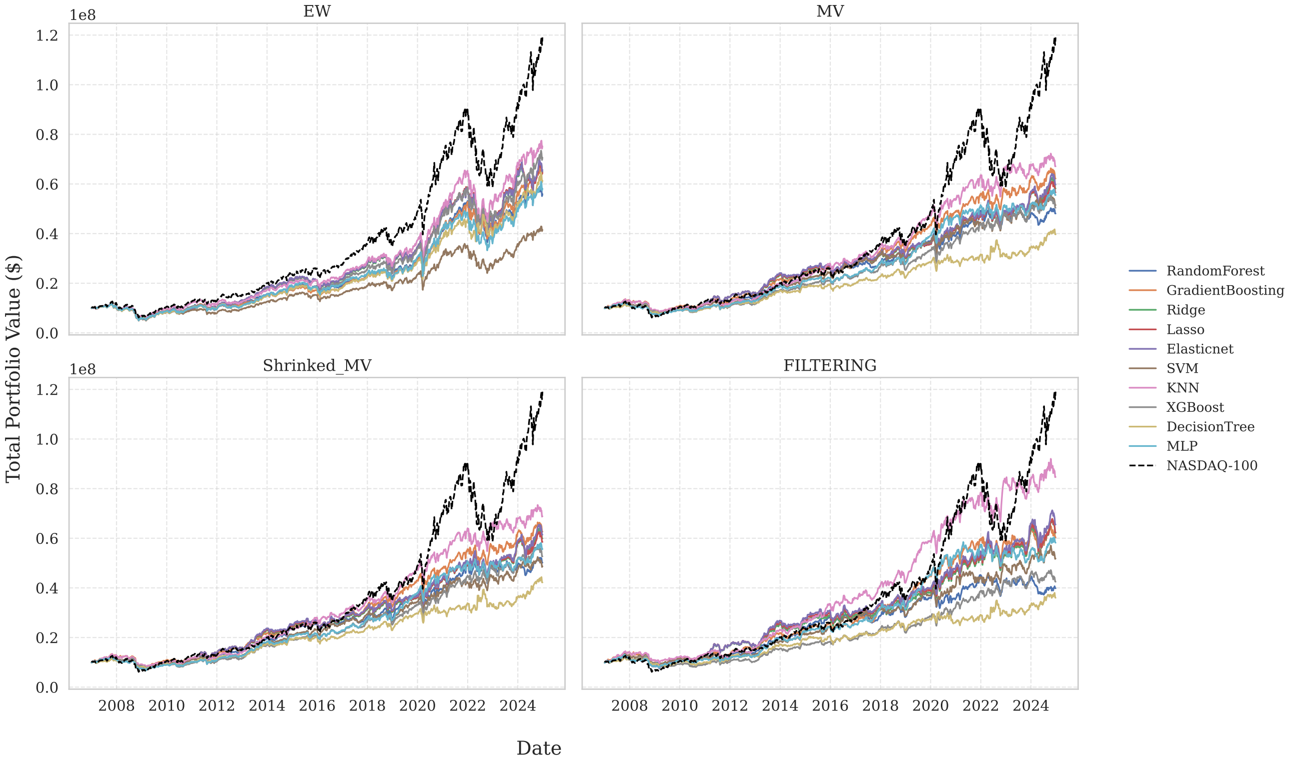

Figure 7: Cumulative returns of individual classifier portfolios by allocation strategy *Each subplot shows the evolution of the portfolio value over time for one of the four allocation methods: Equal Weight (EW), Mean-Variance Optimization (MV), Shrinked MV, and Filtering. All individual models are included in each allocation scheme. Results highlight that no individual model under any allocation strategy is able to consistently outperform the Nasdaq-100 benchmark (black dashed line), further reinforcing the need for ensemble-based forecasts to extract financial value. To reminder threshold techniques without any subscripts take the accuracy metric as reference.*

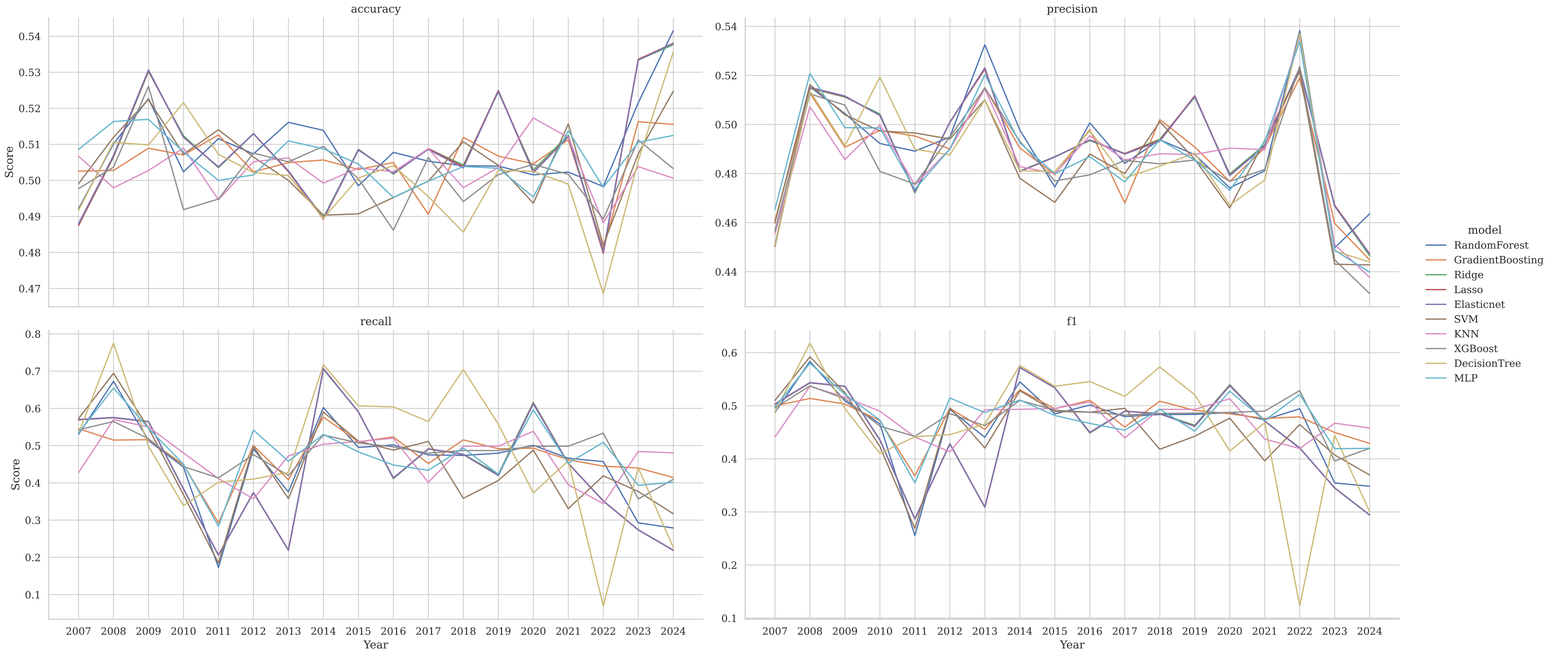

Figure 8: Yearly evolution of classification metrics. *The figure shows the yearly out-of-sample performance of each model for four key metrics: accuracy, precision, recall, and F1-score. Each line corresponds to a specific model, allowing a visual assessment of its temporal behavior. Individual models exhibit fluctuations over time, reflecting the challenges of maintaining consistent classification performance in financial time series.*

Robustness tests

Hyperparameters

Portfolio	Ensemble Method	Threshold	SR	APY [%]	Vol. [%]	DD [%]	VaR [%]	CVaR	WR [%]
FILT.	**T-SA-CV**	QUANTILE	**0.88**	**14.03**	**15.21**	**-33.72**	**-3.12**	**-4.83**	**58.85**
FILT.	T-SA-CV	NAIVE	0.87	13.99	15.23	-33.68	-3.11	-4.83	58.74
FILT.	T-SA-CV	OPTI$_{recall}$	0.87	13.96	15.23	-33.68	-3.11	-4.83	58.74
FILT.	T-SA-CV	CV	0.87	13.94	15.23	-33.72	-3.11	-4.83	58.74
FILT.	T-SA-CV	OPTI$_{PRECISION}$	0.87	13.93	15.23	-33.72	-3.11	-4.83	58.74
FILT.	T-SA-4	CV	0.86	13.74	15.12	-29.75	-3.06	-4.67	58.10
FILT.	T-SA-CV	OPTI	0.86	13.79	15.20	-33.72	-3.13	-4.82	58.85
FILT.	T-SA-4	QUANTILE	0.86	13.67	15.12	-29.75	-3.06	-4.67	58.10
FILT.	T-SA-4	OPTI	0.86	13.67	15.15	-29.75	-3.06	-4.70	58.10
FILT.	T-SA-4	NAIVE	0.86	13.66	15.15	-29.75	-3.06	-4.70	58.10
FILT.	T-SA-4	OPTI$_{recall}$	0.86	13.66	15.14	-29.75	-3.06	-4.70	58.10
FILT.	T-SA-4	OPTI$_{PRECISION}$	0.86	13.65	15.15	-29.71	-3.06	-4.70	58.10
FILT.	T-SA-2	NAIVE	0.85	13.54	15.07	-36.30	-3.13	-4.80	59.17
FILT.	T-SA-2	OPTI$_{recall}$	0.85	13.52	15.07	-36.30	-3.13	-4.80	59.17
FILT.	T-SA-2	QUANTILE	0.85	13.51	15.07	-36.37	-3.13	-4.80	59.17
	INDEX NASDAQ-100		***0.74***	***14.72***	***19.7***	***-51.53***	***-4.66***	***-6.79***	***61.73***

Table 7: Robustness test around hyperparameters optimization. *This table shows the top 15 portfolio rankings when hyperparameter optimization is performed using precision score instead of accuracy.*

In this robustness test, we changed the evaluation metric used for hyperparameters tuning from accuracy to precision. This modification had a significant impact on the model selection and final portfolio rankings. Unlike the previous configuration, most portfolios in the top 15 are now based on trimmed simple average (T-SA) combination methods, with T-SA-CV emerging as the best-performing custom ensemble. This suggests that T-SA strategies are particularly consistent and resilient in financial forecasting tasks, even when the optimization criterion is modified.

Despite these changes, one key element remains stable: the FILT.-based portfolio optimization method (FILT.) consistently dominates the top rankings. This highlights its robustness across different model configurations and confirms its contribution to the overall performance of the strategies.

However, it is important to note that, although some portfolios still outperform the benchmark in terms of Sharpe ratio, none of them surpass it in terms of absolute performance (APY). This result suggests that optimizing models based solely on precision may not fully capture the opportunity cost or the magnitude of return differences, especially in a financial context where ranking errors have asymmetric impacts.

Fees

As expected, the increase in transaction costs has a significant impact on portfolio performance. Given the high frequency of reallocations in our strategy (weekly), these costs accumulate quickly and significantly reduce returns. This robustness test clearly shows that most portfolios underperform the benchmark index and that almost all of them have a negative annualized performance (APY).

These results highlight a major limitation of high-turnover strategies in realistic market conditions. While some models may generate good signals in theory, their profitability can be entirely offset by frictions such as transaction costs. This underscores the need to well incorporate cost considerations into model and portfolio design, particularly in cases of frequent rebalancing.

Portfolio	Ensemble Method	Threshold	SR	APY [%]	Vol. [%]	DD [%]	VaR [%]	CVaR	WR [%]
EW	**CO**	OPTI$_{recall}$	**0.13**	**1.59**	**19.89**	**-56.06**	**-4.67**	**-7.03**	**57.78**
EW	COP_EN	OPTI$_{recall}$	0.07	0.32	19.92	-59.05	-4.78	-7.08	57.78
EW	COP_L2_E	OPTI$_{recall}$	0.02	-0.71	19.90	-55.75	-4.73	-7.06	57.04
EW	COP_L2_IL	OPTI$_{recall}$	-0.05	-2.05	19.93	-59.14	-4.76	-7.08	56.72
EW	COP_CE_E	OPTI$_{recall}$	-0.14	-3.84	19.97	-63.30	-4.83	-7.15	54.80
EW	COP_CE_IL	OPTI$_{recall}$	-0.17	-4.28	19.98	-67.19	-4.80	-7.14	55.01
EW	COS_IL	OPTI$_{recall}$	-0.19	-4.77	19.94	-66.75	-4.76	-7.13	55.44
EW	COP_EN_IL	OPTI$_{recall}$	-0.23	-5.42	19.93	-71.18	-4.83	-7.18	54.80
EW	COP_EN_E	OPTI$_{recall}$	-0.24	-5.63	19.95	-66.79	-4.83	-7.16	54.37
EW	COS_E	OPTI$_{recall}$	-0.26	-6.02	19.93	-74.40	-4.81	-7.16	55.22
EW	COP_L1_IL	OPTI$_{recall}$	-0.30	-6.70	19.89	-76.97	-4.80	-7.19	54.37
Shrinked_MV	CO	OPTI$_{recall}$	-0.32	-3.87	12.79	-54.75	-3.30	-4.44	51.60
MV	CO	OPTI$_{recall}$	-0.38	-4.51	12.79	-60.26	-3.23	-4.41	50.85
EW	COP_L1_E	OPTI$_{recall}$	-0.38	-8.21	19.85	-79.20	-4.83	-7.18	53.30
Shrinked_MV	COP_EN	OPTI$_{recall}$	-0.43	-5.20	12.91	-64.79	-3.34	-4.54	50.53
INDEX NASDAQ-100			***0.74***	***14.72***	***19.7***	***-51.53***	***-4.66***	***-6.79***	***61.73***

Table 8: Robustness test around fees. *This table shows the top 15 portfolio rankings when fees are changed and constantly is 1% of traded value.*

Time period

Portfolio	Ensemble Method	Threshold	SR	APY [%]	Vol. [%]	DD [%]	VaR [%]	CVaR	WR [%]
FILT.	**T-SA-3**	QUANTILE	**0.96**	**13.34**	**12.86**	**-20.75**	**-2.50**	**-4.16**	**43.18**
FILT.	T-SA-3	CV	0.96	13.32	12.86	-20.75	-2.50	-4.15	43.18
FILT.	T-SA-3	OPTI$_{recall}$	0.96	13.31	12.86	-20.75	-2.50	-4.16	43.18
FILT.	T-SA-3	NAIVE	0.96	13.31	12.86	-20.75	-2.50	-4.16	43.18
FILT.	T-SA-3	OPTI	0.96	13.31	12.86	-20.75	-2.50	-4.16	43.18
FILT.	T-SA-3	OPTI$_{PRECISION}$	0.96	13.28	12.86	-20.75	-2.50	-4.15	43.18
FILT.	T-SA-CV	QUANTILE	0.92	12.57	12.65	-19.24	-2.51	-4.09	42.75
FILT.	T-SA-CV	OPTI	0.92	12.56	12.65	-19.24	-2.51	-4.09	42.75
FILT.	T-SA-CV	CV	0.92	12.55	12.65	-19.24	-2.51	-4.09	42.75
FILT.	T-SA-CV	NAIVE	0.92	12.55	12.65	-19.24	-2.51	-4.09	42.75
FILT.	T-SA-CV	OPTI$_{recall}$	0.92	12.55	12.65	-19.24	-2.51	-4.09	42.75
FILT.	T-SA-CV	OPTI$_{PRECISION}$	0.92	12.53	12.65	-19.24	-2.51	-4.09	42.75
FILT.	COP_L1_E	OPTI$_{recall}$	0.91	12.39	12.56	-21.12	-2.71	-4.20	43.60
Shrinked_MV	T-SA-CV	OPTI	0.91	11.67	11.86	-19.04	-2.43	-4.06	44.46
Shrinked_MV	T-SA-CV	QUANTILE	0.91	11.67	11.86	-19.04	-2.43	-4.06	44.56
INDEX NASDAQ-100			***0.64***	***12.48***	***19.98***	***-35.24***	***-4.94***	***-7.15***	***62.02***

Table 9: Robustness test with 2010–2022 period. *This table shows the top 15 portfolio rankings when the time period is adjusted to 2010–2022. The benchmark index is also shifted accordingly.*

A final robustness test is performed to assess how results change when the period under consideration is modified. Naturally, overall returns and volatilities are influenced by specific market conditions during the new period. However, the main objective is to observe whether the best-performing portfolios remain associated with similar optimization approaches and threshold choices.

Once again, we find that the best-performing portfolios are systematically constructed using the Filtering (FILT.) optimization method, confirming its robustness across all periods. In addition, ensemble models based on combinations of truncated averages continue to dominate the rankings. With regard to threshold selection methods, the results appear to be more diverse among the top 15. It should be noted that precision-based threshold optimization is no longer overrepresented, suggesting that its previous dominance may have been specific to the initial period.

Dynamic Clustering in Multi-Factor Copulas with Hidden Markov Models

Abstract

This study proposes a dynamic multi-factor copula model with time-varying, data-driven group assignments. Transitions of firms between groups are modelled using a hidden Markov model, driven by the distance between clusters and the past likelihood of group membership. Using daily returns from S&P 100 stocks between 2015 and 2024, the model is evaluated against two static benchmarks: k-means clustering and industry-based classifications. It was found that the dynamic clustering approach consistently outperforms the static alternatives. Notably, a model with 15 dynamic groups yields better forecasts than an otherwise identical model with 21 static groups. The results show that time-varying group assignments enable the model to adapt to changes in firm characteristics while preserving sufficient persistence in cluster assignments.

1 Introduction

Correlations between asset returns tend to increase significantly during periods of market stress or economic shocks (Chesnay & Jondeau, 2001). This was evident during the 2007–2008 global financial crisis, where models that ignored joint extreme events contributed to the collapse of the housing market (Coval et al., 2009; Zimmer, 2012). This crisis emphasized the need for models that can capture dynamic dependencies in unstable economic conditions. However, financial markets often involve more than 50 variables, resulting in high model complexity (Manner & Reznikova, 2012). This estimation difficulty is reduced by copulas, which separate each variable's marginal behaviour from their dependence structure (Smith, 2015).

Previous research on copulas focused on modelling time-varying dependence through dynamic factor loadings. These loadings represent how strongly each variable is influenced by one or more underlying latent factors, which capture common sources of variation in returns. Oh & Patton (2017) extended this approach by introducing multi-factor copulas with pre-specified static clusters based on SIC industry codes. More recently, Oh & Patton (2023) derived clusters directly from the data using k-means clustering. Assuming stable group assignments over time, they showed that a model with just five data-driven clusters outperforms a comparable model using 21 industry-based clusters in out-of-sample forecasts.

However, no study yet has combined multi-factor copulas with time-varying, data-driven group assignments. Allowing estimated clusters to change over time may better reflect real-world dynamics, such as firms shifting industries, changing strategies or making acquisitions. This idea is supported by João et al. (2023), who developed a linear panel model incorporating a hidden Markov process that allows firms to switch clusters. Their results show that enabling these switches leads to improved model fit. This study investigates whether incorporating time-varying cluster assignments within multi-factor copula models similarly improves predictive performance. Therefore, the research question is: ”How does incorporating time-varying cluster assignments in a high-dimensional multi-factor copula model affect predictive performance relative to static clusters based on industry classifications or k-means clustering?”

To answer this research question, initial clusters were estimated using the k-means algorithm, with group transitions over time modelled via a hidden Markov model. This dynamic clustering approach was compared to two benchmark methods: clustering based on Standard Industrial Classification codes and static k-means clustering. In an empirical analysis, the proposed model was applied to daily returns from stocks in the S&P 100 index over the period 2015–2024 to evaluate its real-world performance. Several copula types were considered, including Gaussian, t, and skewed t, along with both static and dynamic factor loadings.

2 Data and Methodology

2.1 Data

The empirical analysis investigates the daily returns of constituents in the S&P 100 index. The sample period is January 2, 2015, to December 31, 2024, including $T = 2515$ trading days. The dataset contains $N = 98$ stocks that were included in the index as of December 31, 2024, and that were continuously traded throughout the full sample period. A list of all included firms can be found in Appendix A.

2.2 A Dynamic Multi-Factor Copula Model

This study builds on the skewed-t copula model proposed by Oh & Patton (2023). Their approach uses a multi-factor copula model with G clusters of variables, each with a market and group-specific factor loading. In addition, it includes a skewness parameter to capture the asymmetric dependence patterns frequently observed in asset returns. The model is estimated in two stages. First, univariate marginal distributions are fitted to each time series using an AR(1) process for the conditional mean and a GJR-GARCH(1,1) model of Glosten et al. (1993) for the conditional variance. In the second stage of the estimation, conditional on these marginals from the previous stage, the joint dependence among the transformed variables is modelled using a skewed t copula. Time variation in the copula is captured by modelling the factor loadings with Generalized Autoregressive Score (GAS) dynamics (Creal et al., 2013), which updates parameters based on the gradient of the conditional copula log-likelihood.

2.3 Time-Varying Clusters with Markov-switching

The multi-factor copula model assumes G clusters of firms. Clusters can be pre-assigned using industry classifications. Alternatively, Oh & Patton (2023) estimated them from the data using k-means clustering, obtaining static groups fixed over the entire sample period. This study extends their method by modelling cluster memberships as Markov states. Firms can then switch clusters over time, while temporal dependence is preserved (Frühwirth-Schnatter, 2011). The implementation of this method is based on João et al. (2023), who combine a hidden Markov model (HMM) with a linear panel model. Here, their approach is adapted for compatibility with a copula model.

Specifically, the cluster membership of firm i is described by the latent process γ_{it}, where $\gamma_{it} = g$ if firm i belongs to cluster g at time t. The initial cluster assignments are set equal to those obtained by static k-means clustering, as described by Oh & Patton (2023). Let $\pi_{gkt} := \mathbb{P}\{\gamma_{i,t+1} = k \mid \gamma_{it} = g\}$ denote the probability of transitioning from state g to state k at time t. These transition probabilities are assumed to be homogeneous across firms and are collected in the transition matrix Π_t. Assuming transitions are more likely between nearby clusters, transition probabilities π_{gkt} are modelled as a function of distance between clusters:

$$\pi_{gkt} = \frac{\exp(-\delta d_{gk,t-1})}{\sum_{q=1}^{G} \exp(-\delta d_{gq,t-1})} \qquad g, k = 1, \dots, G, \tag{1}$$

where $d_{gk,t}$ denotes the distance between clusters g and k at time t and $\delta \geq 0$ is a parameter that controls how fast the transition probability decays as the distance increases. Distances are based on the common market factor:

$$d_{gk,t} =\mid \lambda^M_{g,t} - \lambda^M_{k,t} \mid . \tag{2}$$

Since cluster memberships are unobserved, they are inferred using filtered probabilities. These probabilities, denoted by $\tau_{ig,t|t} := \mathbb{P}[\gamma_{it} = g \mid \mathcal{F}_t; \theta]$, contain the probability that unit i belongs to cluster g at time t, conditional on the observed data up to time t, $\mathcal{F}_t$. The copula likelihood at time t is then computed by summing over the likelihood of all possible cluster states, weighted by their predicted probabilities $\tau_{ig,t|t-1}$. Note that the copula defines a joint distribution over all units simultaneously and thus does not allow decomposition into individual likelihoods. To address this, the conditional mixture likelihood is defined (DeSarbo & Cron, 1988). This likelihood accounts for uncertainty in the cluster assignment of a single unit i, while keeping the cluster assignments of all other units fixed at their previously estimated values:

$$\mathbf{c}^{(i)}(\mathbf{u}_t \mid \hat{\Gamma}_{t-1}, \mathcal{F}_{t-1}; \boldsymbol{\theta}) = \sum_{g=1}^{G} \tau_{ig,t|t-1} \cdot \mathbf{c}(\mathbf{u}_t \mid \tilde{\Gamma}_{i,g,t-1}, \mathcal{F}_{t-1}; \boldsymbol{\theta}), \tag{3}$$

where $\mathbf{c}(\cdot)$ is the likelihood of the static Gaussian copula, $\hat{\Gamma}_t$ is the vector of estimated cluster assignments at time t and $\tilde{\Gamma}_{i,g,t}$ is identical to $\hat{\Gamma}_t$ except that variable i is reassigned to cluster g.

The predicted cluster probabilities $\tau_{ig,t+1|t}$ are updated recursively using the forward algorithm (Hamilton, 1989). The Markov property implies that the probability that firm i is in cluster g at time $t+1$ depends on the probability that it is currently in cluster k, and on the probability of transitioning from cluster k to g:

$$\tau_{ig,t+1|t} = \mathbb{P}[\gamma_{i,t+1} = g \mid \mathcal{F}_t; \boldsymbol{\theta}] = \sum_{k=1}^{G} \pi_{kgt} \mathbb{P}[\gamma_{it} = k \mid \mathcal{F}_t; \boldsymbol{\theta}] = \sum_{k=1}^{G} \pi_{kgt} \tau_{ik,t|t}. \tag{4}$$

The second step of the forward algorithm is updating the filtered probabilities $\tau_{ig,t|t}$ via Bayes' rule. The predicted probabilities are combined with the conditional likelihood of the observed data under each potential cluster assignment for firm i, while keeping the cluster assignments of all other firms fixed:

$$\begin{aligned} \tau_{ig,t|t} = \mathbb{P}[\gamma_{it} = g \mid \hat{\Gamma}_{t-1}, \mathcal{F}_t; \boldsymbol{\theta}] &= \frac{\tau_{ig,t|t-1}\ \mathbf{c}(\mathbf{u}_t \mid \tilde{\Gamma}_{i,g,t-1}, \mathcal{F}_{t-1}; \boldsymbol{\theta})}{\mathbf{c}^{(i)}(\mathbf{u}_t \mid \hat{\Gamma}_{t-1}, \mathcal{F}_{t-1}; \boldsymbol{\theta})} \\ = \frac{\tau_{ig,t|t-1}\ \mathbf{c}(\mathbf{u}_t \mid \tilde{\Gamma}_{i,g,t-1}, \mathcal{F}_{t-1}; \boldsymbol{\theta})}{\tau_{i1,t|t-1}\ \mathbf{c}(\mathbf{u}_t \mid \tilde{\Gamma}_{i,1,t-1}, \mathcal{F}_{t-1}; \boldsymbol{\theta}) + \cdots + \tau_{iG,t|t-1}\mathbf{c}(\mathbf{u}_t \mid \tilde{\Gamma}_{i,G,t-1}, \mathcal{F}_{t-1}; \boldsymbol{\theta})}&. \end{aligned} \tag{5}$$

The forward filtering algorithm described in Eqs.(4)-(5) is performed for each variable, after which each unit is assigned to the cluster with the highest posterior probability:

$$\gamma_{i,t} = \arg\max_g \tau_{ig,t|t}. \tag{6}$$

For robustness, the forward filtering algorithm is run repeatedly for all firms, updating their cluster assignments until no firm switches clusters between iterations. In practice, convergence is typically achieved after just one iteration.

Regarding the out-of-sample forecasts, each variable is assigned to the cluster with the highest predicted probability from Eq.(4), after which the probabilities are updated using Bayes's rule for use in the next time period.

2.4 Model Estimation and Evaluation

Joint estimation of the dynamic copula parameters and time-varying group assignments is computationally demanding. Therefore, a three-stage estimation procedure is used. A description of this procedure is provided in Appendix B. After estimation, model performance is assessed using the Akaike Information Criterion, the economic relevance of the clusters, and forecast accuracy.

Forecast accuracy is assessed using two scoring rules: the log-likelihood and the conditional likelihood score. The log-likelihood scoring rule of Amisano & Giacomini (2007) measures predictive performance over the full support of copula model M_i, and is defined as

$$S_{l,t}(\hat{\mathbf{u}}_t, M_i) = \log c_t(\hat{\mathbf{u}}_t \mid \hat{\theta}_{C,t}, M_i), \tag{7}$$

where $c_t(\cdot)$ denotes the copula density at time t, and $\hat{\mathbf{u}}_t$ is the vector of estimated uniform marginals. Dependence in the joint lower tail is especially important in risk management and finance, as extreme losses often occur simultaneously. Therefore, the copula is evaluated using the conditional likelihood score proposed by Diks et al. (2014), which focuses on the lower tail:

$$S_{cl,t}(\hat{\mathbf{u}}_t, M_i) = \Big(\log c_t(\hat{\mathbf{u}}_t \mid \hat{\theta}_{C,t}, M_i) - \log C_t(\mathbf{q} \mid \hat{\theta}_{C,t}, M_i)\Big) \times I[\hat{\mathbf{u}}_t < \mathbf{q}], \tag{8}$$

where $\mathbf{q}$ is an $N \times 1$ threshold vector, $C_t(\cdot \mid \hat{\theta}_{C,t}, M_i)$ is the copula distribution, and $I[\hat{\mathbf{u}}_t < \mathbf{q}] = \prod_{i=1}^{N} I[\hat{u}_{i,t} < q_i]$ indicates joint threshold exceedance. Equation (8) thus measures the log-likelihood of model M_i conditional on the event $\hat{\mathbf{u}}_t < \mathbf{q}$, corresponding to the lower tail region $[0, q_1] \times \cdots \times [0, q_N]$. To allow for time variation, the threshold vector is defined as $\mathbf{q}_t = (\bar{q}_t, \ldots, \bar{q}_t)$, where $\bar{q}_t$ satisfies

$$\frac{1}{1000} \sum_{j=1}^{1000} I[\hat{\mathbf{u}}_{t-j} < \mathbf{q}_t] = q,$$

for a specified tail probability q, such as 0.05. To compare predictive accuracy across models while controlling for multiple testing, the Diebold & Mariano (2002) test is combined with the Model Confidence Set procedure of Hansen et al. (2011), which iteratively eliminates the worst-performing models until a subset of statistically indistinguishable models remains.

3 Results

3.1 Marginal Model

The marginal model is estimated for each individual return series, with detailed results reported Table 3 in Appendix C. Summary statistics show heavy tails and slight negative skewness in the returns (see Figure 1 in Appendix C), while standardized residuals retain mild skewness and substantial excess kurtosis, supporting the skewed t model. Heterogeneity in pairwise correlations further motivates a copula model to capture non-linear and asymmetric dependencies across stock returns.

3.2 In-sample Evaluation of Dynamic and Benchmark Clusters

Using transformed residuals from the marginal model, the copula model is estimated to capture dependence between return series. Time-varying clusters are benchmarked against k-means and industry-based clusters (one- and two-digit Standard Industry Classification (SIC) codes), using a Gaussian copula with static factor loadings. Model fit is evaluated with the Akaike Information Criterion (AIC), where lower values indicate better performance. The optimal value of the transition decay parameter δ (see Eq. (1)) is found by grid search to be 20.

The results show that time-varying clusters achieve significantly better AIC values than static clustering methods, confirming the potential of data-driven dynamic group assignments as suggested by João et al. (2023). Figure 2 in Appendix D shows AIC values across different numbers of clusters G, indicating that 21 is the optimal number of groups for both static and dynamic clusters. Additionally, a model with only six estimated groups outperforms the 21 two-digit SIC groups, consistent with the findings of Oh & Patton (2023).

3.3 Estimated Cluster Assignments

To investigate the economic relevance of cluster transitions, the initial static clusters are compared with the clusters at the end of the sample period. A complete overview of these clusters is provided in Table 4 and Table 5 in Appendix D, respectively. Factor loadings for the model with dynamic cluster assignments, estimated for Gaussian, t, and skewed t copulas with GAS dynamics, are reported in Table 6. The static groups show strong alignment with two-digit SIC classifications and generally show meaningful coherence, although some inconsistencies remain. Allowing for time-varying transitions via the hidden Markov model resolves several of these inconsistencies. For example, General Electric, initially clustered with financial firms, is reassigned to group 4 with industrial companies such as General Dynamics, better reflecting its core business. Improvements are also seen in within-group correlations, with Figure 3 in Appendix D showing that time-varying clusters yield higher correlations for General Electric's group and two other cases.

Across the full sample, 183 transitions are observed, with around 4% of stocks switching each quarter. Figure 4 and Figure 5 in Appendix E report the fraction of stocks switching each quarter and the number of transitions per stock, respectively. Transition rates are low early in the sample but rise from 2020, peaking in early 2021 during the COVID-19 crisis. Elevated rates are also observed in the second and fourth quarters of 2024, possibly linked to geopolitical and economic uncertainty. About 42.9% of firms never change clusters. A small number of stocks switch repeatedly ("flickering"), likely reflecting proximity to multiple cluster boundaries. Some groups do not experience changes, such as groups 12 and 13 (energy and utilities, and oil and gas), which may be related to the steady and clearly defined nature of these industries.

3.4 Out-of-sample Forecast Performance

Next, the dynamic model is evaluated out-of-sample using a rolling estimation window of 1,000 observations, resulting in 1,515 out-of-sample observations, ranging from December 21, 2018 to December 31, 2024. Model parameters are re-estimated every 250 observations, and a one-step-ahead copula density forecast is constructed for each day. For the dynamic clustering model, the transition parameter is set to the optimal in-sample value $\delta = 20$, which also demonstrated robust performance in the out-of-sample evaluation.

Table 1 presents the results of the copula density forecast evaluation across different group sizes for SIC, static k-means, and dynamic clustering. It reports the time-averaged log-likelihood scores $S_{l,t}$, 5% left-tail conditional log-likelihood scores $S_{cl,t}$, and p-values from the Model Confidence Set (MCS). The left panel shows results for static factor loadings, and the right panel for GAS dynamic loadings. The table shows three interesting results. First, dynamic clustering consistently outperforms static clustering across all group sizes under the log scoring rule, in line with in-sample results. In both panels, the model with 21 dynamic groups achieves the highest average log-likelihood, with an average log-likelihood of 31.44 under static loadings and 31.94 with GAS dynamics. In addition, both 21-group models achieve an MCS p-value of 1.00, indicating statistically superior predictive performance relative to all other models. Second, dynamic clustering improves the 5% conditional log score relative to the static clusters, suggesting better performance in capturing joint downside risk. However, the improvement is smaller than for the full log-likelihood, and the MCS includes some static models. Third, allowing for time-varying factor loadings through GAS dynamics leads to further improvements in both the overall and conditional left-tail log-likelihood, even when combined with dynamic clustering.

Table 1
One-step ahead copula density forecasts

	Static loadings		GAS loadings	
	Full	5% tail	Full	5% tail
	$S_{l,t}$(p-val)	$S_{cl,t}$(p-val)	$S_{l,t}$(p-val)	$S_{cl,t}$(p-val)
SIC 1-digit static	24.80(0.00)	2.395(0.00)	24.89(0.00)	2.297(0.00)
SIC 2-digit static	27.48(0.00)	2.424(0.00)	27.62(0.00)	2.289(0.00)
3 static groups	24.14(0.00)	2.401(0.00)	24.47(0.00)	2.363(0.00)
6 static groups	26.80(0.00)	2.417(0.00)	27.24(0.00)	2.413(0.00)
9 static groups	27.86(0.00)	2.417(0.00)	28.14(0.00)	2.415(0.00)
12 static groups	28.88(0.00)	2.422(0.01)	29.29(0.00)	2.414(0.00)
15 static groups	28.96(0.00)	**2.443(0.07)**	29.38(0.00)	2.428(0.00)
18 static groups	30.00(0.00)	**2.447(0.14)**	30.54(0.00)	2.439(0.00)
21 static groups	30.24(0.00)	**2.453(0.31)**	30.59(0.00)	2.456(0.00)
24 static groups	28.62(0.00)	2.334(0.03)	28.97(0.00)	2.438(0.00)
SIC 1-digit dynamic	27.40(0.00)	2.401(0.00)	27.93(0.00)	2.374(0.00)
SIC 2-digit dynamic	29.95(0.00)	2.436(0.00)	30.22(0.00)	2.418(0.00)
3 dynamic groups	24.93(0.00)	2.289(0.00)	25.58(0.00)	2.405(0.00)
6 dynamic groups	27.48(0.00)	2.348(0.00)	27.95(0.00)	2.417(0.00)
9 dynamic groups	28.14(0.00)	2.412(0.00)	28.62(0.00)	2.418(0.00)
12 dynamic groups	29.06(0.00)	2.419(0.00)	29.43(0.00)	2.421(0.00)
15 dynamic groups	30.30(0.00)	**2.451(0.28)**	30.76(0.00)	**4.497(0.08)**
18 dynamic groups	30.56(0.00)	**2.453(0.31)**	31.02(0.00)	**2.509(0.72)**
21 dynamic groups	**31.44(1.00)**	**2.459(0.76)**	**31.94(1.00)**	**2.516(1.00)**
24 dynamic groups	29.68(0.00)	**2.464(1.00)**	29.89(0.00)	2.483(0.00)

Note: This table reports the accuracy of one-step-ahead copula density forecasts for daily returns of S&P 100 stocks, using a multi-factor copula model with Student's t distribution and transition decay $\delta = 20$. The mean log score ($S_{l,t}$) and the 5% conditional log-likelihood score ($S_{cl,t}$) for the lower tail are shown. p-values from the Model Confidence Set (MCS) procedure of Hansen et al. (2011) are reported in parentheses, with bold numbers indicating models that belong to the MCS of their column at a significance level of 5%. The out-of-sample period runs from December 21, 2018, to December 31, 2024, including 1,515 observations. Note that the 1-digit SIC clusters consist of 8 groups and the 2-digit SIC clusters contain 21 groups.

To further evaluate out-of-sample forecasting performance, alternative copula specifications were compared, considering both static versus dynamic (GAS) factor loadings and different copula families (Gaussian, t and skew t). The significance of differences in out-of-sample likelihoods is assessed using the Diebold & Mariano (2002) test, with standard errors computed with the Newey & West (1987) estimator based on 10 lags. All DM-test results can be found in Table 7 and 8 in Appendix F for static and dynamic clustering, respectively.

Three main findings were obtained. First, all test statistics comparing static and GAS versions of the HMM models were found to be positive and highly significant, indicating that the incorporation of GAS dynamics improves the fit across all copula types and group sizes. The effect was observed to be particularly strong for the skewed t copula relative to the Gaussian and Student's t copulas. Second, when copula families were compared under GAS dynamics, the Student's t copula was consistently found to outperform both the Gaussian and skewed t copulas across all group sizes. This outcome is consistent with the findings of Oh & Patton (2023). Third, the skewed t copula was shown to perform substantially worse in out-of-sample forecasts, in some cases performing even worse than the Gaussian copula. The poor performance of the skewed t copula may be due to the inherent penalty that forecast evaluations impose on estimation uncertainty. Unless additional parameters differ significantly from zero and are estimated with sufficient accuracy, the model may achieve better predictive accuracy by excluding them altogether. Finally, Table 7 shows that, in general, the patterns observed for dynamic clusters also hold for static clusters. However, while the skewed t copula still performs poorly, it is not consistently outperformed by the Gaussian copula across all group sizes.

3.5 Economic determinants of Forecast Performance

Forecast performance is further evaluated across economic environments using the Conditional Equal Predictive Ability (CEPA) and Conditional Superior Predictive Ability (CSPA) tests. The CEPA test of Giacomini & White (2006) examines whether predictive accuracy depends on market conditions, while the CSPA test proposed by Li et al. (2022) assesses whether any alternative model systematically outperforms the benchmark. Economic conditions are summarized by market volatility (VIX), cross-sectional dispersion of returns, and the alpha from the Capital Asset Pricing Model (CAPM).

The detailed results of both tests are shown in Table 9 in Appendix G. The results indicate that the dynamic clustering model consistently outperforms static benchmarks. Compared to industry-based clusters, gains are not strongly linked to economic variables. In contrast, when compared to k-means clusters, improvements in forecasting performance are larger in periods of high volatility, high dispersion, and greater CAPM alpha, with alpha being most influential. The CSPA test confirms that dynamic clustering dominates static benchmarks across the entire range of conditions.

4 Conclusion

This paper extends the factor copula model of Oh & Patton (2023) by relaxing the assumption of static group assignments. Firms often adjust their strategy, enter or exit markets, and engage in mergers or acquisitions. As a result, the dependence structures among firms may also change over time, making fixed clusters less realistic. Therefore, this study proposes a copula model that incorporates time-varying clusters using a hidden Markov model, adapting the approach developed for multivariate panel data by João et al. (2023).

In the empirical application to daily returns of S&P 100 stocks from 2015 to 2024, the dynamic clustering model consistently outperforms static benchmarks based on SIC codes and k-means clustering. The improvement appears to be driven by firms that undergo changes in their business activities or were initially misclassified. These gains are evident in both in-sample and out-of-sample evaluations and are particularly strong during periods of high volatility, elevated dispersion, and large CAPM alpha values. However, the model introduces additional computational demands, not all clusters are economically interpretable and some firms switch frequently between clusters.

Future research could address the issue of frequent switching between clusters by incorporating non-Markovian transitions, which prevent switches if a stock has recently switched. Alternatively, flickering could be reduced by using the non-parametric method of João et al. (2024), which uses a modified version of k-means clustering to ensure temporal stability in clusters. A challenge of the latter is its high computational cost, as it requires running the k-means algorithm for each point in time. Another potential extension is to explore whether including additional explanatory variables improves model performance. Variables such as market capitalization or return on equity could be used to inform the transition probabilities between clusters, an approach shown to enhance performance by João et al. (2023). Lastly, future work could focus on scaling the model to accommodate larger datasets containing more firms or other asset classes.

References

Amisano, G. & Giacomini, R. (2007). Comparing density forecasts via weighted likelihood ratio tests. *Journal of Business & Economic Statistics*, *25*(2), 177-190.

Chesnay, F. & Jondeau, E. (2001). Does correlation between stock returns really increase during turbulent periods? *Economic Notes*, *30*(1), 53-80.

Coval, J., Jurek, J. & Stafford, E. (2009). The economics of structured finance. *Journal of Economic Perspectives*, *23*(1), 3-25.

Creal, D. D., Koopman, S. J. & Lucas, A. (2013). Generalized autoregressive score models with applications. *Journal of Applied Econometrics*, *28*(5), 777-795.

DeSarbo, W. S. & Cron, W. L. (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of classification*, *5*, 249-282.

Diebold, F. X. & Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & economic statistics*, *20*(1), 134-144.

Diks, C., Panchenko, V., Sokolinskiy, O. & van Dijk, D. (2014). Comparing the accuracy of multivariate density forecasts in selected regions of the copula support. *Journal of Economic Dynamics and Control*, *48*, 79-94.

Frühwirth-Schnatter, S. (2011). Panel data analysis: a survey on model-based clustering of time series. *Advances in Data Analysis and Classification*, *5*, 251-280.

Giacomini, R. & White, H. (2006). Tests of conditional predictive ability. *Econometrica*, *74*(6), 1545-1578.

Glosten, L. R., Jagannathan, R. & Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *The journal of finance*, *48*(5), 1779-1801.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica: Journal of the econometric society*, 357-384.

Hansen, P. R., Lunde, A. & Nason, J. M. (2011). The model confidence set. *Econometrica*, *79*(2), 453-497.

João, I. C., Lucas, A., Schaumburg, J. & Schwaab, B. (2023). Dynamic clustering of multivariate panel data. *Journal of Econometrics*, *237*(2), 105281.

João, I. C., Schaumburg, J., Lucas, A. & Schwaab, B. (2024). Dynamic nonparametric clustering of multivariate panel data. *Journal of Financial Econometrics*, *22*(2), 335-374.

Li, J., Liao, Z. & Quaedvlieg, R. (2022). Conditional superior predictive ability. *The Review of Economic Studies*, *89*(2), 843-875.

Manner, H. & Reznikova, O. (2012). A survey on time-varying copulas: specification, simulations, and application. *Econometric reviews*, *31*(6), 654-687.

Newey, W. K. & West, K. D. (1987). Hypothesis testing with efficient method of moments estimation. *International Economic Review*, 777-787.

Oh, D. H. & Patton, A. J. (2017). Modeling dependence in high dimensions with factor copulas. *Journal of Business & Economic Statistics*, *35*(1), 139-154.

Oh, D. H. & Patton, A. J. (2023). Dynamic factor copula models with estimated cluster assignments. *Journal of Econometrics*, *237*(2), 105374.

Smith, M. S. (2015). Copula modelling of dependence in multivariate time series. *International Journal of Forecasting*, *31*(3), 815-833. (ID: 271676)

Zimmer, D. M. (2012). The role of copulas in the housing crisis. *Review of Economics and Statistics*, *94*(2), 607-620.

Appendix A
List of Firms Used in the Empirical Analysis

Table 2
Summary of firms in the S&P 100

Ticker	Name	SIC	Ticker	Name	SIC	Ticker	Name	SIC
AAPL	Apple	35	DHR	Danaher	50	MS	Morgan Stanley	62
ABBV	Abbvie	28	DIS	Disney	73	MSFT	Microsoft	73
ABT	Abbott Lab.	50	DUK	Duke Energy	49	NEE	Nextera Energy	49
ACN	Accenture	67	EMR	Emerson	35	NFLX	Netflix	78
ADBE	Adobe	73	F	Ford	37	NKE	Nike	30
AIG	Ame Inter	63	FDX	Fedex	45	NVDA	Nvidia	36
AMD	Adv Micro Dev	36	GD	Gen Dynamics	37	ORCL	Oracle	73
AMGN	Amgen	28	GE	Gen Electric	35	PEP	Pepsico	20
AMT	American Tower	48	GILD	Gilead	28	PFE	Pfizer	28
AMZN	Amazon	73	GM	General Motors	37	PG	Procter Gamble	28
AVGO	Broadcom	36	GOOG	Alphabet	73	PM	Philip Morris	21
AXP	Amex	60	GOOGL	Alphabet	73	QCOM	Qualcomm	36
BA	Boeing	37	GS	Goldman Sachs	62	RTX	RTX	37
BAC	Bank of Am	60	HD	Home Depot	52	SBUX	Starbucks	58
BH	Biglari Holdings	58	HON	Honeywell Int	50	SCHW	Schwab Charles	62
BK	Bank of NY	60	IBM	IBM	73	SO	Southern	49
BKNG	Booking	73	INTC	Intel	36	SPG	Simon Property	67
BLK	Blackrock	62	INTU	Intuit	73	T	AT&T	48
BMY	Bristol-Myers	28	JNJ	Johnson&J	28	TGT	Target	53
C	Citigroup	60	JPM	Jpmorgan	60	TMO	Thermo Fisher	38
CAT	Caterpillar	35	KO	Coca Cola	20	TMUS	T-Mobile	48
CHTR	Charter Comm	48	LLY	Lilly Eli	28	TSLA	Tesla	37
CL	Colgate Palmo	28	LMT	Lockheed Mar	37	TXN	Texas Instru	36
CMCSA	Comcast	48	LOW	Lowes	52	UNH	Unitedhealth	63
COF	Capital One	60	MA	Mastercard	73	UNP	Union Pacific	40
COP	Conocophillips	13	MCD	Mcdonalds	58	UPS	United Parcel	45
COST	Costco	53	MDLZ	Mondelez Int	20	USB	US Bancorp	60
CRM	Salesforce	73	MDT	Medtronic	38	V	Visa	73
CSCO	Cisco Sys	36	MET	Metlife	63	VZ	Verizon	48
CVS	C V S Health	59	META	Meta	73	WFC	Wells Fargo	60
CVX	Chevron	13	MMM	3M	50	WMT	Walmart	53
D	Dominion En	49	MO	Altria Group	21	XOM	Exxon Mobil	29
DE	Deere	35	MRK	Merck	28			

SIC	Description	Num	SIC	Description	Num	SIC	Description	Num
1	Mining, construct.	2	4	Transprt, comm's	13	7	Services	15
2	Manuf: food, furn.	16	5	Trade	13	9	Non-classifiable	2
3	Manuf: elec, mach	20	6	Finance, Ins	17	Total		98

Note: This table reports the ticker symbol, company name, and the first two digits of the SIC code for the firms included in the empirical analysis. The sample consists of firms that were constituents of the S&P 100 index as of December 31, 2024, and that were continuously traded throughout the full sample period (2015–2024). Firms that underwent mergers or splits during the sample period are excluded. SIC codes correspond to those assigned as of the midpoint of the sample period (December 31, 2019). For firms that changed their ticker symbol or name, the most recent identifiers are reported.

Appendix B
Estimation procedure

Due to the computational complexity of jointly estimating dynamic copula parameters and time-varying group assignments, a three-stage estimation procedure is adopted. In the first stage, the initial cluster assignments and copula parameters are estimated following Oh & Patton (2023). Static group assignments $\hat{\Gamma}_1$ are obtained by applying k-means clustering to the full sample using a misspecified static Gaussian copula. Conditional on these initial clusters, the copula parameters $\boldsymbol{\psi} = [\omega_1^M, \ldots, \omega_G^M, \omega_1^C, \ldots, \omega_G^C, \alpha^M, \beta^M, \alpha^C, \beta^C, \nu, \zeta]'$ are estimated by maximizing the log-likelihood of the skewed t copula:

$$\hat{\boldsymbol{\psi}} = \arg\max_{\boldsymbol{\psi}} \sum_{t=1}^{T} \log \mathbf{c}_{\text{Skew }t,t}(\mathbf{u}_t; \boldsymbol{\psi} \mid \hat{\Gamma}_1). \tag{9}$$

The skewed t copula nests both the Gaussian and symmetric t copulas as special cases. In both the simulation study and the empirical analysis, all three copulas are estimated for comparison.

In the second stage, time-varying cluster assignments are estimated. For each time period t, the filtered probabilities $\tau_{ig,t|t}$ are updated using the forward algorithm in Eqs. (4)-(5), again based on the misspecified static Gaussian copula. Each stock is then assigned to the cluster with the highest posterior probability, yielding the current group assignment $\hat{\Gamma}_t$.

In the third stage, the dynamic factors loadings are updated using the Generalized Autoregressive Score dynamics, based on the copula parameters obtained in the first stage. GAS adjusts the loadings in response to new information by leveraging the gradient of the log-likelihood function of the conditional copula. The update equations take the form:

$$\lambda_{g,t+1}^M = \omega_g^M + \alpha^M \frac{\partial \log \mathbf{c}_{\text{Skew}t,t}(\mathbf{x}_t; \mathbf{R}_t, \nu, \zeta)}{\partial \lambda_{g,t}^M} + \beta^M \lambda_{g,t}^M, \quad \text{for } g = 1, \ldots, G \tag{10}$$

$$\lambda_{g,t+1}^C = \omega_g^C + \alpha^C \frac{\partial \log \mathbf{c}_{\text{Skew}t,t}(\mathbf{x}_t; \mathbf{R}_t, \nu, \zeta)}{\partial \lambda_{g,t}^C} + \beta^C \lambda_{g,t}^C, \quad \text{for } g = 1, \ldots, G \tag{11}$$

where $\mathbf{x}_t = T_{\text{skew}}^{-1}(\mathbf{u}_t; \nu, \zeta)$, and $\mathbf{c}_{\text{Skew}t,t}(\mathbf{x}_t; \mathbf{R}_t, \nu, \zeta)$ denotes the conditional skewed t copula density. Finally, given the current factor loadings and group assignments, the log-likelihood of each observation is computed.

With respect to the remaining parameters, the number of clusters G must be chosen. Although the Akaike Information Criterion (AIC) is commonly used, it may overestimate the number of clusters (Frühwirth-Schnatter, 2011). Therefore, the optimal number of clusters is also validated based on out-of-sample forecasting performance. In theory, the number of clusters could vary over time if all firms were to transition out of a given cluster, implying $G = G_{max}$. However, to preview the results, this has not occurred in practice, likely because the initial static clustering provides a sufficiently accurate starting point. Moreover, maintaining a fixed number of clusters over time is common in literature and appears to be a reasonable assumption (Frühwirth-Schnatter, 2011; João et al., 2023). The decay parameter δ in Eq. (1), which controls the dynamics of cluster transitions, is selected by evaluating model performance over a grid of values.

Appendix C
Summary Statistics and Marginal Model Results

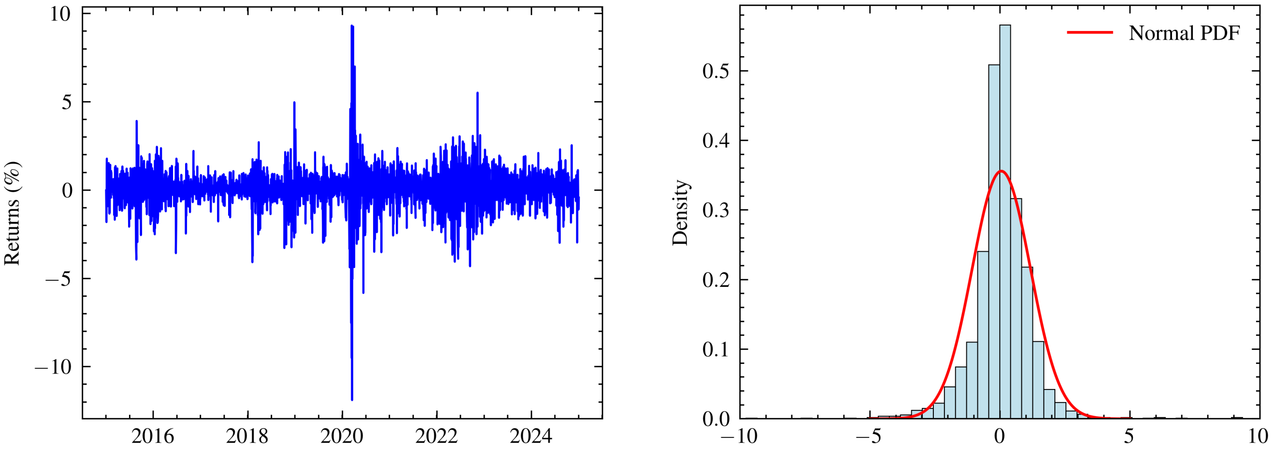

Figure 1. Distribution of S&P 100 returns. The left panel displays the daily value-weighted market returns over the full sample period from 2015 to 2024. The right panel shows the distribution of the returns, along with a fitted Gaussian distribution.

Table 3
Summary statistics for the marginal model

	Cross-sectional distribution					
	Mean	5%	25%	Median	75%	95%
Panel A: Marginal moments						
Mean	0.001	0.000	0.000	0.001	0.001	0.001
Std	0.018	0.012	0.015	0.017	0.020	0.026
Skewness	-0.028	-0.786	-0.228	0.079	0.237	0.819
Kurtosis	15.000	8.201	10.676	13.352	17.703	28.429
Panel B: Marginal model parameters						
Constant	0.001	0.000	0.000	0.001	0.001	0.001
AR(1)	-0.020	-0.052	-0.038	-0.019	-0.004	0.014
$\varpi\times10^4$	0.009	0.002	0.004	0.006	0.011	0.023
α	0.034	0.001	0.016	0.028	0.043	0.085
κ	0.091	0.016	0.062	0.089	0.119	0.161
β	0.893	0.799	0.871	0.897	0.929	0.957
ξ	4.558	3.555	3.950	4.415	4.941	6.128
ψ	-0.035	-0.092	-0.063	-0.036	-0.010	0.025
Panel C: Correlations of standardized residuals						
Pearson	0.288	0.127	0.214	0.274	0.347	0.484
Spearman	0.330	0.152	0.251	0.316	0.398	0.535

Note: This table reports the cross-sectional distribution of summary statistics from 98 daily return series spanning January 2, 2015 to December 31, 2024. Panel A reports the distribution of the first four moments of the returns, Panel B shows the estimated parameters from the marginal models, and Panel C provides a summary of the pairwise correlations among the standardized residuals.

Appendix D
Comparison between Static and Dynamic Clusters

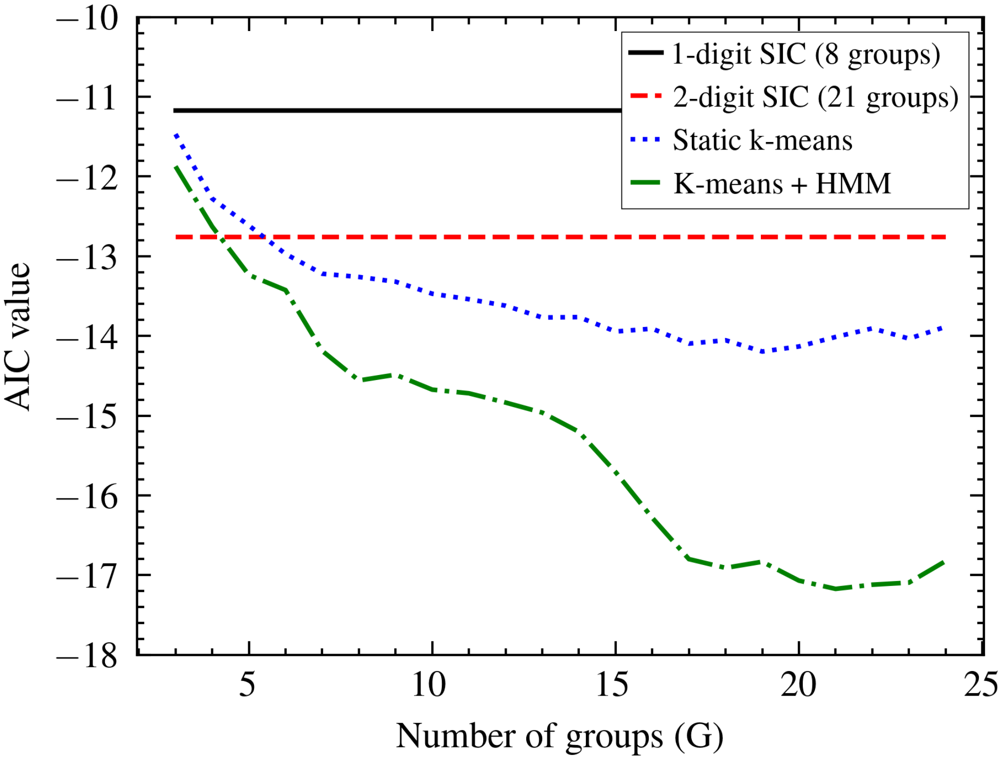

Figure 2. AIC values as a function of the number of groups (G) for static clustering and dynamic clustering combining k-means with a hidden Markov model. For comparison, AIC values corresponding to the 1-digit and 2-digit SIC-based groupings, comprising 8 and 21 groups, respectively, are also shown. Lower AIC values indicate a better model fit. The y-axis is scaled by 10^{-4}.

Table 4
Estimated group assignments with static clustering

Group	Ticker	Name	SIC	Group	Ticker	Name	SIC
1	ABBV	Abbvie	28	8	CAT	Caterpillar	35
	ABT	Abbott Lab.	28		EMR	Emerson	36
	AMGN	Amgen	28		FDX	Fedex	45
	BMY	Bristol-Myers	28		UNP	Union Pacific	40
	GILD	Gilead	28		UPS	United Parcel	45
	JNJ	Johnson&J	28				
	LLY	Lilly Eli	28	9	AMZN	Amazon	73
	MDT	Medtronic	38		BKNG	Booking	73
	MRK	Merck	28		META	Meta	73
	PFE	Pfizer	28		NFLX	Netflix	78
	TMO	Thermo Fisher	38				
	UNH	Unitedhealth	63	10	ACN	Accenture	73
					CSCO	Cisco Sys	36
2	BAC	Bank of Am	62		IBM	IBM	73
	BK	Bank of NY	60		ORCL	Oracle	73
	C	Citigroup	62				
	COF	Capital One	60	11	COST	Costco	53
	GS	Goldman Sachs	62		CVS	C V S Health	59
	JPM	Jpmorgan	60		TGT	Target	53
	MET	Metlife	63		WMT	WalMart	53
	MS	Morgan Stanley	60				
	SCHW	Schwab Charles	62	12	D	Dominion En	49
	USB	US Bancorp	60		DUK	Duke Energy	49
	WFC	Wells Fargo	60		NEE	Nextera Energy	49
					SO	Southern	49
3	AMT	American Tower	67				
	CL	Colgate Palmo	28				
	KO	Coca Cola	20	13	COP	Conocophillips	13
	MDLZ	Mondelez Int	20		CVX	Chevron	29
	MO	Altria Group	21		XOM	Exxon Mobil	29
	PEP	Pepsico	20				
	PG	Procter Gamble	28	14	AIG	Ame Inter	63
	PM	Philip Morris	21		AXP	Amex	61
	SPG	Simon Property	67		GE	Gen Electric	35
4	BA	Boeing	37	15	BH	Biglari Holdings	58
	GD	Gen Dynamics	37		TMUS	T-Mobile	48
	HON	Honeywell Int	37		TSLA	Tesla	37
	LMT	Lockheed Mar	37				
	MMM	3M	38	16	MCD	Mcdonalds	58
	RTX	RTX	37		NKE	Nike	30
					SBUX	Starbucks	58
5	ADBE	Adobe	73				
	CRM	Salesforce	73	17	F	Ford	37
	INTU	Intuit	73		GM	General Motors	37
	MA	Mastercard	73				
	MSFT	Microsoft	73	18	GOOG	Google	73
	V	Visa	73		GOOGL	Google	73
6	AAPL	Apple	35	19	DE	Deere	35
	AMD	Adv Micro Dev	36		INTC	Intel	36
	AVGO	Broadcom	36				
	NVDA	Nvidia	36	20	BLK	Blackrock	62
	QCOM	Qualcomm	36		DHR	Danaher	38
	TXN	Texas Instru	36				
				21	HD	Home Depot	52
7	CHTR	Charter Comm	48		LOW	Lowes	52
	CMCSA	Comcast	48				
	DIS	Disney	48				
	T	AT&T	48				
	VZ	Verizon	48				

Note: This table reports the estimated static group assignments for 21 groups, as determined by the AIC-selected optimal number of clusters. For firms that have changed their ticker symbol or name, the most recent identifiers are shown. Groups are ordered by the number of stocks.

Table 5
Estimated final group assignments with dynamic clusters

Group	Ticker	Name	SIC	Group	Ticker	Name	SIC
1	ABBV	Abbvie	28	9	AAPL	Apple	35
	ABT	Abbott Lab.	28		AMZN	Amazon	73
	AMGN	Amgen	28		COST	Costco	53
	BMY	Bristol-Myers	28		LLY	Lilly Eli	28
	GILD	Gilead	28		META	Meta	73
	JNJ	Johnson&J	28		NFLX	Netflix	78
	KO	Coca Cola	20				
	MRK	Merck	28	10	ACN	Accenture	73
	PEP	Pepsico	20		CSCO	Cisco Sys	36
	PFE	Pfizer	28		HON	Honeywell Int	37
					IBM	IBM	73
2	BAC	Bank of Am	62		MMM	3M	38
	BK	Bank of NY	60				
	C	Citigroup	62	11	CHTR	Charter Comm	48
	COF	Capital One	60		CMCSA	Comcast	48
	GS	Goldman Sachs	62		CVS	C V S Health	59
	JPM	JPMorgan	60		MDLZ	Mondelez Int	20
	MET	Metlife	63		MDT	Medtronic	38
	MS	Morgan Stanley	60		UNH	Unitedhealth	63
	USB	US Bancorp	60				
	WFC	Wells Fargo	60	12	D	Dominion En	49
					DUK	Duke Energy	49
3	AMT	American Tower	62		NEE	Nextera Energy	49
	CL	Colgate Palmo	60		SO	Southern	49
	DIS	Disney	62				
	PG	Proctor Gamble	60	13	COP	Conocophilips	13
	PM	Philip Morris	62		CVX	Chevron	29
	SPG	Simon Property	60		XOM	Exxon Mobil	29
	T	AT&T	63				
	VZ	Verizon	60	14	AIG	Ame Inter	63
	WMT	Walmart	60		AXP	Amex	61
					BA	Boeing	37
4	GD	Gen Dynamics	37				
	GE	Gen Electric	35	15	BH	Biglari Holdings	58
	LMT	Lockheed Mar	37		TMUS	T-Mobile	48
	RTX	RTX	37		TSLA	Tesla	37
5	ADBE	Adobe	73	16	BKNG	Booking	73
	CRM	Salesforce	73		NKE	Nike	30
	INTU	Intuit	73		PCLN	Priceline	73
	MSFT	Microsoft	73		SBUX	Starbucks	58
	ORCL	Oracle	73				
				17	F	Ford	37
6	AMD	Adv Micro Dev	36		GM	General Motors	37
	AVGO	Broadcom	36				
	INTC	Intel	36	18	GOOG	Google	73
	NVDA	Nvidia	36		GOOGL	Google	73
	QCOM	Qualcomm	36				
	TXN	Texas Instru	36	19	DE	Deere	35
					MCD	Mcdonalds	58
7	DHR	Danaher	38				
	TGT	Target	53	20	BLK	Blackrock	62
	TMO	Thermo Fisher	38		MA	Mastercard	73
					SCHW	Schwab Charles	62
8	CAT	Caterpillar	35		UNP	Union Pacific	40
	EMR	Emerson	36		V	Visa	73
	FDX	Fedex	45				
	UPS	United Parcel	45	21	HD	Home Depot	52
					LOW	Lowes	52

Note: This table reports the estimated dynamic group assignments for 21 groups at the end of the sample period. For firms that have changed their ticker symbol or name, the most recent identifiers are shown. Cluster numbers are consistent with the original numbering assigned in the static clustering.

Table 6
Estimation results for the optimal 21 group model

Panel A: Parameter estimation accuracy						
	Gaussian		t		Skew t	
	est.	s.e.	est.	s.e.	est.	s.e.
ω_1^M	0.052	0.003	0.007	0.001	0.006	0.001
ω_2^M	0.105	0.007	0.016	0.003	0.013	0.001
ω_3^M	0.053	0.003	0.007	0.001	0.006	0.001
ω_4^M	0.074	0.005	0.011	0.001	0.009	0.001
ω_5^M	0.083	0.005	0.012	0.002	0.008	0.001
ω_6^M	0.072	0.005	0.010	0.000	0.009	0.001
ω_7^M	0.055	0.004	0.008	0.001	0.007	0.001
ω_8^M	0.081	0.005	0.012	0.001	0.010	0.001
ω_9^M	0.061	0.004	0.009	0.001	0.007	0.001
ω_{10}^M	0.078	0.005	0.011	0.001	0.009	0.001
ω_{11}^M	0.053	0.003	0.008	0.001	0.006	0.001
ω_{12}^M	0.050	0.003	0.007	0.002	0.006	0.001
ω_{13}^M	0.084	0.005	0.012	0.001	0.010	0.001
ω_{14}^M	0.073	0.005	0.011	0.001	0.009	0.001
ω_{15}^M	0.039	0.002	0.005	0.001	0.005	0.001
ω_{16}^M	0.064	0.004	0.009	0.002	0.008	0.001
ω_{17}^M	0.105	0.007	0.016	0.001	0.013	0.002
ω_{18}^M	0.520	0.032	0.076	0.002	0.063	0.008
ω_{19}^M	0.063	0.004	0.011	0.002	0.009	0.001
ω_{20}^M	0.078	0.005	0.009	0.001	0.014	0.002
ω_{21}^M	0.118	0.007	0.017	0.003	0.008	0.001
ω_1^C	0.003	0.001	0.002	0.001	0.002	0.000
ω_2^C	0.005	0.001	0.002	0.001	0.004	0.001
ω_3^C	0.003	0.001	0.003	0.001	0.003	0.000
ω_4^C	0.003	0.001	0.002	0.001	0.002	0.000
ω_5^C	0.003	0.001	0.003	0.001	0.003	0.000
ω_6^C	0.003	0.001	0.003	0.001	0.003	0.000
ω_7^C	0.002	0.001	0.002	0.001	0.002	0.000
ω_8^C	0.002	0.001	0.002	0.001	0.002	0.000
ω_9^C	0.002	0.001	0.003	0.001	0.003	0.000
ω_{10}^C	0.001	0.000	0.002	0.001	0.002	0.000

Table 6
Estimation results for the optimal 21 group model, continued

Panel A: Parameter estimation accuracy

	Gaussian		t		Skew t	
	est.	s.e.	est.	s.e.	est.	s.e.
ω_{11}^{C}	0.002	0.001	0.002	0.001	0.002	0.000
ω_{12}^{C}	0.007	0.001	0.006	0.002	0.006	0.001
ω_{13}^{C}	0.007	0.001	0.007	0.002	0.007	0.001
ω_{14}^{C}	0.002	0.001	0.002	0.001	0.002	0.000
ω_{15}^{C}	0.001	0.001	0.001	0.001	0.000	0.000
ω_{16}^{C}	0.002	0.001	0.002	0.001	0.002	0.000
ω_{17}^{C}	0.006	0.001	0.005	0.001	0.001	0.001
ω_{18}^{C}	0.037	0.006	0.034	0.007	0.034	0.004
ω_{19}^{C}	0.002	0.001	0.001	0.001	0.001	0.001
ω_{20}^{C}	0.000	0.000	0.000	0.000	0.001	0.001
ω_{21}^{C}	0.007	0.001	0.006	0.002	0.006	0.002
α^{M}	0.033	0.001	0.011	0.003	0.010	0.001
β^{M}	0.917	0.005	0.987	0.008	0.990	0.002
α^{C}	0.006	0.001	0.008	0.002	0.008	0.001
β^{C}	0.995	0.005	0.995	0.008	0.996	0.001
ν			0.034	0.003	0.034	0.001
ζ					-0.398	0.001

Panel B: Estimation details

	Gaussian	t	Skew t
$\log \mathcal{L}$	86643.05	89266.23	87968.92
AIC	-173194	-178438	-175841
BIC	-172926	-178164	-175562
Time (clustering) (hrs)	1.23	1.23	1.23
Time (copula) (hrs)	2.71	2.98	2.68
EM iterations	95.54	95.54	95.54

Note: This table reports the estimated parameters and standard errors for the multi-factor copula model with dynamic group assignments, estimated via a Hidden Markov Model. Results are presented for the Gaussian, t, and skew-t copulas. The model was estimated on the full sample using 21 groups, selected as optimal based on the Akaike Information Criterion (AIC). Panel B reports model fit measures, computational time, and the number of EM iterations. All estimations were performed on a machine with an Apple M1 processor (8 cores).

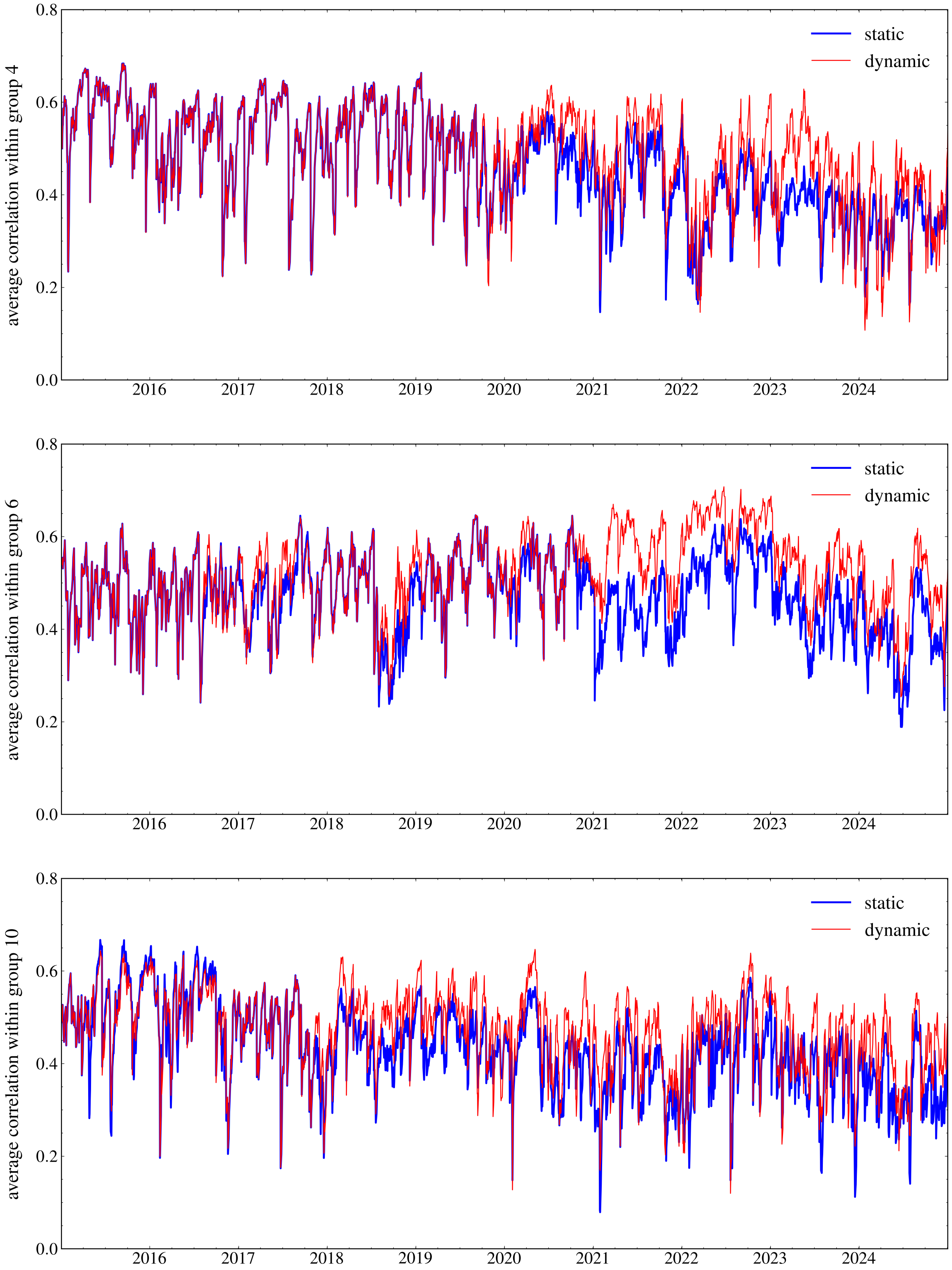

Figure 3. Model-implied rank correlations over the full sample period. The upper panel displays correlations for group 4 (from Table 4), comparing static clusters with dynamic clusters from the Markov-switching model. The middle panel and lower shows the same comparison for group 6 and 10, respectively. The results are obtained from the model with GAS dynamics and the Gaussian copula.

Appendix E
Transition Dynamics

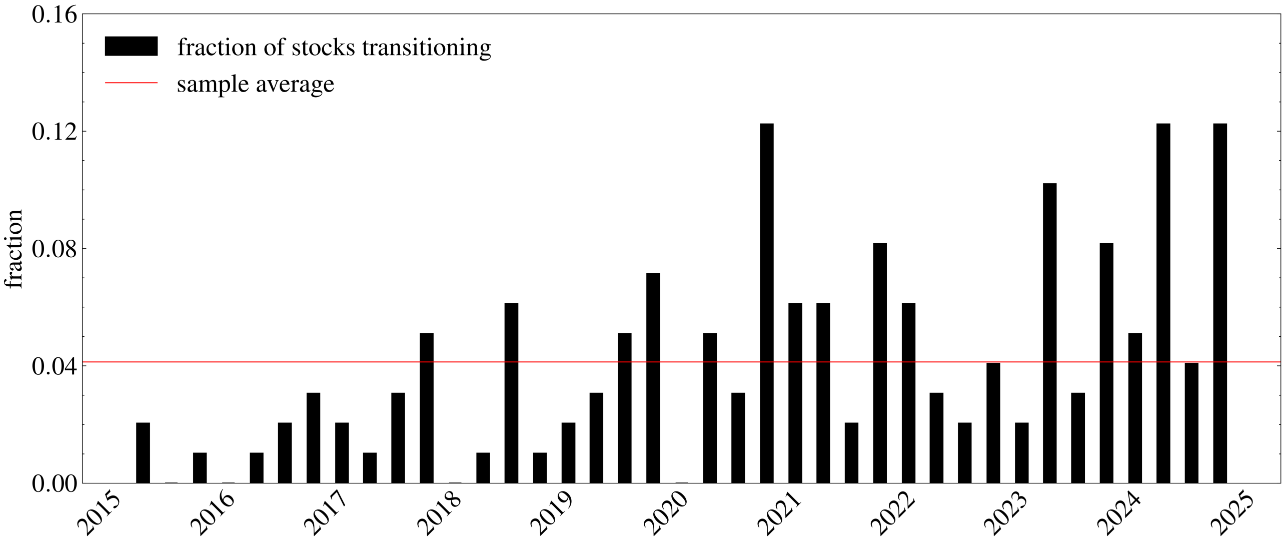

Figure 4. Timing of cluster transitions. The black bars indicate the fraction of stocks that are estimated to change groups for each quarter between 2015Q1 and 2024Q4. The red line reports the average transition frequency over the full sample.

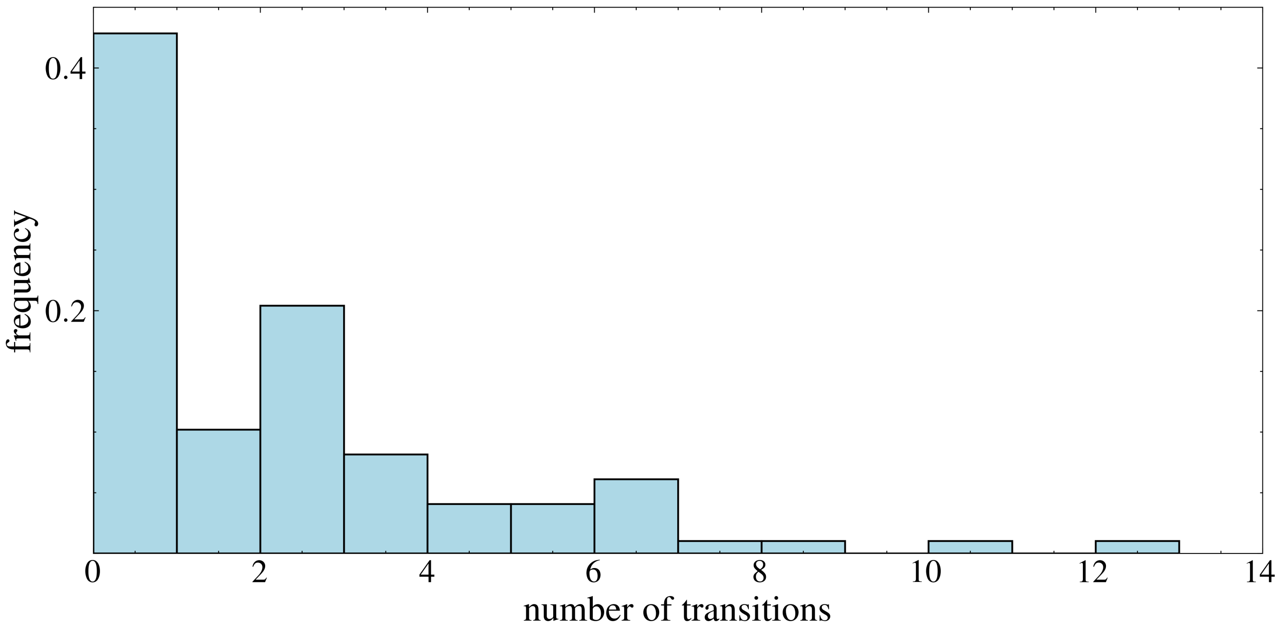

Figure 5. Histogram of cluster transitions counts per stock.

Appendix F
Diebold-Mariano test results

Table 7
Comparison of different copulas for static clustering

	Static vs. GAS			Copula shape		
	Gaussian	t	skew$t$	G vs. t	G vs. skew$t$	t vs. skew$t$
SIC 1-digit	3.170	3.704	27.771	12.223	-3.025	-18.945
SIC 2-digit	1.905	3.728	31.834	12.748	10.320	-8.435
3 groups	0.327	1.632	24.721	11.212	9.305	-9.1612
6 groups	3.017	4.294	30.483	12.466	11.088	-6.035
9 groups	3.807	5.172	7.481	12.741	6.647	-17.572
12 groups	3.574	5.117	11.443	13.057	7.085	-17.954
15 groups	3.058	4.528	22.474	12.276	-5.912	-17.475
18 groups	5.412	6.968	34.708	12.238	10.625	-7.451
21 groups	4.218	6.680	35.023	12.582	9.942	-7.051
24 groups	3.532	6.140	20.083	12.996	-5.367	-20.002

Note: This table reports Diebold-Mariano t statistics for pairwise comparisons of models with static clusters using their out-of-sample log-likelihood. The left panel compares models assuming static factor loadings with those using GAS dynamics, for a Gaussian, t and skew-t copula and for a variety of choices for the number of groups. The right panel compares the different copula shapes, using GAS dynamics in all cases, across a variety of choices for the number of groups. In a comparison labelled “A vs. B,” a positive t-statistic implies that model B outperforms model A, whereas a negative t-statistic suggests that model A performs better than model B.

Table 8
Comparison of different copula’s for dynamic clustering

	Static vs. GAS			Copula shape		
	Gaussian	t	skew$t$	G vs. t	G vs. skew$t$	t vs. skew$t$
SIC 1-digit	11.497	11.404	30.732	11.858	-8.489	-19.712
SIC 2-digit	5.601	5.605	21.014	12.843	-7.388	-18.094
3 groups	5.129	4.852	21.643	11.879	-5.861	-18.594
6 groups	5.833	5.249	18.869	12.299	-7.487	-19.124
9 groups	6.058	5.424	13.251	12.900	-9.863	-14.459
12 groups	5.121	4.140	20.386	13.246	-6.182	-18.543
15 groups	4.974	4.618	25.893	12.363	-6.521	-17.847
18 groups	5.841	6.795	32.697	12.290	-7.342	-17.624
21 groups	6.767	7.261	31.105	12.906	-6.699	-21.958
24 groups	5.489	6.434	29.453	11.358	-7.231	-21.762

Note: This table reports Diebold-Mariano t statistics for pairwise comparisons of models with dynamic HMM clusters using their out-of-sample log-likelihood. The left panel compares models assuming static factor loadings with those using GAS dynamics, for a Gaussian, t and skew-t copula and for a variety of choices for the number of groups. The right panel compares the different copula shapes, using GAS dynamics in all cases, across a variety of choices for the number of groups. In a comparison labeled “A vs. B,” a positive t-statistic implies that model B outperforms model A, whereas a negative t-statistic suggests that model A performs better than model B.

Appendix G
Giocomini & White and CSPA test results

Table 9
Economic determinants of forecast performance

	Dynamic vs. SIC 2-digit				Dynamic vs. static k-means			
Intercept	1.716	1.716	1.716	1.716	1.349	1.349	1.349	1.349
(s.e.)	(0.107)	(0.107)	(0.107)	(0.106)	(0.112)	(0.111)	(0.103)	(0.104)
[t-stat]	[16.106]	[16.106]	[16.058]	[16.181]	[12.094]	[12.182]	[13.103]	[12.925]
VIX	-0.056			-0.142	0.226			-0.016
(s.e.)	(0.082)			(0.099)	(0.104)			(0.124)
[t-stat]	[-0.683]			[-1.437]	[2.168]			[-1.039]
Dispersion		0.057		0.235		0.657		0.328
(s.e.)		0.090		(0.126)		(0.155)		(0.130)
[t-stat]		[0.631]		[1.872]		[4.243]		[2.528]
Abs. alpha			-0.096	-0.188			0.895	0.750
(s.e.)			(0.092)	(0.105)			(0.118)	(0.124)
[t-stat]			[-1.038]	[-1.791]			[7.589]	[6.043]
R^2 (%)	0.030	0.030	0.086	0.368	0.393	3.325	6.164	6.599
GW p-value$_{\text{ALL}}$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
GW p-value$_{\text{SLOPES}}$	0.495	0.528	0.299	0.403	0.000	0.000	0.000	0.000
CSPA p-value	0.000	0.000	0.000	-	0.000	0.000	0.000	-

Note: This table reports the results of the Giacomini & White (2006) tests and the Li et al. (2022) (CSPA) tests. The benchmark model is the model with 21 time-varying clusters with GAS dynamics and a student t copula. The conditioning variables are the VIX index, the dispersion (cross-sectional standard deviation of returns) and the absolute value of the cross-sectional average CAPM alpha. For the GW tests, conditioning variables are standardized to ensure the comparability of the test statistics.

Multiscale Inefficiency Index

August 11, 2025

Abstract

This paper investigates the long-term memory and multifractal properties of financial time series through Hurst exponent estimation techniques, including both the classical R/S statistic and its modified version (M-R/S). Recognizing the limitations of a single static Hurst exponent often distorted by trends, sample length, or structural breaks we complement our analysis with Multifractal Detrended Fluctuation Analysis (MF-DFA), which reveals the local scaling dynamics and multifractal spectrum of the data. Building on these insights, we introduce a novel inefficiency index that integrates two key dimensions: the width of the multifractal spectrum, capturing scale-invariant long-range correlations, and the deviation of a rolling Hurst exponent from the efficient market benchmark of 0.5, indicating momentum or mean-reversion. To showcase the practical value of our index, we design a long/short trading strategy that uses it to filter out false signals from the Hurst exponent, thereby improving performance compared to a traditional long/short Hurst-based approach.

1 Introduction

The Hurst exponent is a crucial tool for analyzing long-term memory and self-similarity in stochastic processes. Originally introduced by Harold Hurst in the 1950s for studying river flows, this measure has since been widely adopted in various fields such as physics, environmental science and finance. In financial markets, the Hurst exponent serves as an indicator to determine whether a time series exhibits long-range dependence (a value greater than 0.5) or mean-reverting behavior (a value less than 0.5), a value equals to 0.5 indicates that the series follows a pure random walk, characteristic of standard Brownian motion.

The most common method for estimating the Hurst exponent is through Rescaled Range (R/S) analysis, introduced by Hurst and later refined by Mandelbrot. However, the traditional R/S statistic has its limitations, particularly its sensitivity to short-term memory effects, which can obscure the detection of long-term memory. To mitigate these issues, **lo1991** proposed a modified version of the R/S statistic (M-R/S) that better accounts for short-term autocorrelation.

In this study, we apply both the R/S method and the M-R/S to estimate the Hurst exponent on financial time series and we complement our analysis with Multifractal Detrended Fluctuation Analysis (MF-DFA), which examines the local behavior of the series and characterizes its multifractal spectrum. That way we can capture the local scaling dynamics and identify the presence of multifractality, which is often indicative of complex market behaviors possibly inefficient.

The Fractional Brownian motion (fBm) is often used as a benchmark model for processes with memory, as it embodies the scaling properties and persistence typically observed in long-memory data. While fBm provides a theoretical framework for understanding these phenomena, our study focuses on practical estimation methods.

2 Literature Review

Memory diagnostics in finance revolve around R/S and its modified M-RS tests; Mandelbrot and Wallis pioneered the use of the rescaled range (R/S) statistic to detect long-term memory in geophysical and financial time series, highlighting its sensitivity to persistent and anti-persistent behaviors **mandelbrot1968**; **mandelbrot1969a**; **mandelbrot1969b**. Mandelbrot further emphasized the limitations of classical methods and the need for robust estimators in the presence of nonstationarity and structural breaks **mandelbrot1973**; **mandelbrot1979**. Moreover, **dimatteo2007** review details their scope and wavelet refinements. Kwapień show that shuffling kills multifractality, proving that inefficiency derived from the width of the spectrum stem from temporal correlations rather than heavy tails **kwapien2023**.

3 Fractional Brownian Motion

Fractional Brownian motion (fBm) is a generalization of standard Brownian motion that introduces dependence in increments, making it suitable for modeling processes with memory effects. It is a continuous-time Gaussian process $X_H(t)$ where $H \in [0, 1]$ corresponds to the Hurst exponent with the following properties:

— The process exhibits self-similarity, meaning that for any scaling factor c, $c \in \mathbb{R}^+$, the rescaled process satisfies:

$$X_H(ct) \overset{d}{=} c^H X_H(t). \tag{1}$$

where the symbol $\overset{d}{=}$ denotes equality in distribution, meaning that the statistical properties of $X_H(ct)$ and $c^H X_H(t)$ are identical.

- The increments $X_H(t) - X_H(s)$ follow a normal distribution with mean zero and variance :

$$\mathbb{E}\left[(X_H(t) - X_H(s))^2\right] = \sigma^2 |t-s|^{2H}, \tag{2}$$

where H is the Hurst exponent.

- When $H = 0.5$, fBm reduces to classical Brownian motion.
- For $H > 0.5$, the process exhibits long-term positive autocorrelation, meaning that an increase in the past tends to be followed by further increases.
- For $H < 0.5$, the process has anti-persistent behavior, where an increase in the past is more likely to be followed by a decrease.

The covariance function of fBm is given by (see Section 7.1 for demonstration):

$$C_H(t,s) = \frac{\sigma^2}{2}\left(t^{2H} + s^{2H} - |t-s|^{2H}\right), \tag{3}$$

which accounts for the dependence structure of the process. The Hurst exponent H plays a critical role in determining the smoothness and correlation properties of fBm:

- **For small H values** ($H < 0.5$), the process is highly erratic, with rapid changes and weak memory effects.
- **For large H values** ($H > 0.5$), the trajectory becomes smoother, and the process exhibits long-range dependence.

3.1 Data

The data used in this analysis is monthly and consists of the historical closing prices of five major stock market indices: the S&P 500, Russell 2000, FTSE 100, Nikkei 225, and the DAX. The data spans the period from September 10th, 1987, to February 28th, 2025.

For each index, the closing price time series was transformed using the natural logarithm to obtain a series of log prices. Additionally, a stationarity test was conducted on the log prices series using the Augmented Dickey-Fuller (ADF) test. The results indicated that all series were non-stationary, suggesting the presence of unit roots. To address this, the log prices were differentiated once, after which they exhibited stationarity (test are available in Table 1).

These differentiated log returns were then used to calculate the R/S and modified R/S statistics and estimate the Hurst exponent. The purpose of using this data is to evaluate the long-term memory properties of financial markets, which can indicate persistence or mean-reversion in market behavior.

3.2 Results

The results of the Hurst exponent estimation using the traditional R/S method are presented in Table 3.

Based on the results obtained from applying the traditional R/S method, all the series appear to exhibit long-term memory, as the Hurst exponents are consistently greater than 0.5. However, the unknown asymptotic distribution of the traditional R/S statistic prevents us from determining whether these Hurst values are statistically significant. To address this, we use the modified R/S method, comparing the statistic V to the critical values provided by **lo1991** (1.620 at the 10% level and 1.747 at the 5% level in a one-tailed test). Our analysis shows that only one series the returns of the Russell 2000 small and mid cap (US) exhibits statistically significant persistence, for the other series, despite Hurst exponents greater than 0.5, the null hypothesis of short memory cannot be rejected.

Since it seems unrealistic to characterize series with only one static Hurst exponent as the series might be influenced by local trends, periods estimations, frequencies and to gain deeper insight into the local scaling dynamics of these series, we now turn to Multifractal Detrended Fluctuation Analysis (MF-DFA). Specifically, MF-DFA allows us to investigate the variety of local behaviors present in the time series by characterizing its multifractal spectrum. This spectrum reveals how "rough" or "smooth" different segments of the series are and indicates the prevalence of each level of irregularity. By examining the multifractal spectrum, we can determine whether the data exhibits a wide range of scaling behaviors, indicative of multifractality, or if it behaves more uniformly. This transition to MF-DFA thus provides a complementary perspective that deepens our understanding of the complex, scale-dependent dynamics governing the indices and how it relates to the Hurst exponent.

3.3 Generalized Hurst Exponent

The S&P 500 and Russell 2000 are ideal candidates for multifractal analysis. The modified R/S statistic (M-R/S) for the S&P 500 is approximately 0.501 very close to 0.5 which suggests that its dynamics are consistent with efficient market behavior. In contrast, the Russell 2000 has a modified Hurst exponent of approximately 0.588, indicating significant long-range dependencies and a less efficient market.

These discrepancies between the two series highlight their distinct scaling properties and market efficiencies. Our aim with the multifractal analysis is to capture and quantify these differences in local scaling behavior. By analyzing the multifractal spectrum of each index, we hope to match these structural discrepancies, thereby providing deeper insights into the dynamics of each market. By analyzing both the S&P 500 and Russell 2000, we gain insight into how differences in efficiency and persistence affect their multifractal characteristics. For this analysis, we will use the daily returns of the Russell 2000 index, S&P 500 index from September 10th, 1987, to February 28th, 2025 (about 10 000 data points).

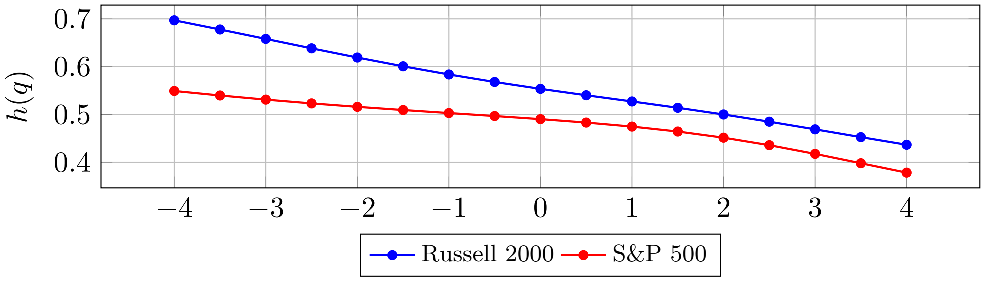

Figure 1 – Generalized Hurst exponent $h(q)$ for the S&P 500 returns. Values of q are equally spaced between -4 and 4. The scale used are logly spaced between 10 and 500.

If we take a closer look at the results from MF-DFA, we observe that the generalized Hurst exponent, $h(q)$, varies as a function of q. The decrease sloping is a sign that the serie exhibits multifractal behavior. In a monofractal process, h(q) remains constant, reflecting uniform scaling. Variation of h(q) with q indicates that small and large fluctuations scale differently. The curvature of the line indicates the presence of heterogeneity in the distribution of singularities, with different regions of the series characterized by varying degrees of irregularity. Lower values of q emphasize small fluctuations, while higher values highlight high fluctuations. Therefore, this spectrum showcases that during periods of small fluctuations (q $<$ 0) the series is likely to exhibit long-term memory as the Hurst exponent is greater than 0.5, whether for drastic changes (q $>$ 0) in the series behavior the Hurst exponent is likely not to be high. This result is consistent with our simulation of the fractional Brownian motion (see 7.4), where we can see that the series exhibits smooth and regular behavior (calm fluctuations) for high Hurst exponent and sharply irregular behavior (high fluctuations) for low Hurst exponent.

The generalized Hurst exponent for the S&P 500 returns exhibits a similar behavior to that of the Russell 2000 returns, except that it is less pronounced. At q = -4, the series exhibits a Hurst exponent of 0.56 compared to 0.7 for the Russell 2000, those series seems to slightly differs in their behavior. This difference, albeit modest, may hint at distinct market microstructure characteristics between the two indices. For instance, the S&P 500, with its larger and more liquid companies, might experience a smoothing effect on return dynamics that could reduce the observable multifractality. In contrast, the Russell 2000, representing smaller-cap stocks, may be subject to greater fluctuations and market inefficiencies, which could amplify multifractal behavior. Overall, our findings provide an interesting perspective on market behavior, suggesting that although both indices share similar multifractal characteristics, subtle variations exist that could reflect underlying market differences.

From the MF-DFA analysis, we can also compute the Hölder exponent and multifractal spectrum. Calculating the Hölder exponent and multifractal spectrum extends MF-DFA by detailing local behavior. This logical continuation deepens insights into the complex, heterogeneous dynamics of the market.

3.4 Hölder exponent

The Hölder exponent $\alpha(q)$ characterizes the local multifractal strength of a signal and is obtained using the Legendre transform of $h(q)$:

$$\alpha(q) = h(q) + qh'(q). \tag{4}$$

where $h'(q)$ is the derivative of $h(q)$ with respect to q. This exponent quantifies the intensity of local singularities: lower values of α indicate highly irregular (or sharply singular) behavior, while higher values correspond to smoother regions of the signal. Thus, the Hölder exponent reveals the heterogeneity of fluctuations within the signal. This exponent describes the degree of multifractal in different parts of the series, revealing the heterogeneity of fluctuations.

3.5 Multifractal Spectrum

The multifractal spectrum $f(\alpha)$ provides a measure of the fractal dimension of subsets characterized by a given α:

$$f(\alpha) = q[\alpha(q) - h(q)] + 1. \tag{5}$$

This spectrum describes the distribution of singularities in the time series. A wider spectrum indicates stronger multifractality.

The analysis using the Hölder exponent and multifractal spectrum is a powerful tool for studying complex systems. In particular, it enables one to identify and quantify regions of strong multifractal, which may correspond to extreme events or sudden changes in dynamics and to describe the distribution and frequency of irregular behaviors in time series. Thus, the multifractal approach offers a detailed and nuanced description of a signal's local variability, providing essential insights for understanding and predicting its underlying dynamics.

We can distinguish two main contributions to the multifractal spectrum:

$$\mathcal{M}(q) \propto \underbrace{f_{\text{tail}}(q)}_{\substack{\text{Strongly non-Gaussian}\\ \text{distribution}}} \quad \textit{and} \quad \underbrace{f_{\text{corr}}(q)}_{\substack{\text{Temporal correlations}\\ \text{in the series}}}$$

Therefore, in the literature, the multifractality is often reffered as two types :

Type I multifractality arises from a broad probability density function of the series values **kantelhardt2002** whereas Type II multifractality stems from long-range correlations within the time series. This distinction enables us to identify and quantify the type of multifractality present. By shuffling the series, we effectively eliminate the long-range correlations, retaining only the influence of the value distribution. By using a phase randomization algorithm we can generate a surrogate which keeps the long term correlation in the series intact and make the distribution gaussian. The difference in width between spectrums showcase the degree of multifractality and hence inefficience of both sources.

The multifractal spectrum $f(\alpha)$ for the Russell 2000 (see Figure 5) returns exhibits a bell-shaped curve, indicating multiple scaling behaviors in the data. Furthermore, the approximate symmetry of the curve around its maximum implies that both large and small fluctuations are represented, albeit with varying intensity. Overall, this bell-shaped spectrum underscores the complex, multiscale nature of the Russell 2000 returns. Comparing it with the shuffled series, we observe that the width of the spectrum is lower than the original series, indicating that the shuffled series exhibits a more uniform behavior with less multifractality. The surrogate version of the series is narrower meaning that most of the multifractality that we observe are due to the non-gaussian distribution of returns. The multifractality linked to the non-gaussian distribution cannot be quantified as true multifractality since it's due to the finite sample size (**kwapien2023**). The only true multifractality comes from the long term correlation therefore, we use the width of the surrogate spectrum to quantifie for it.

See section 8.4 for the S&P 500 returns.

4 Inefficiency Index

4.1 Proposition of an Inefficiency Index

Market inefficiency is captured by two structural components: the width of the multifractal spectrum, $\Delta\alpha = \alpha_{\text{max surrogate}} - \alpha_{\text{min surrogate}}$, and the deviation of the rolling Hurst exponent from the efficient market value, $|H_{\text{rolling}} - 0.5|$. We define the inefficiency index as

$$I = \Delta\alpha_{\text{surrogate}} \times |H_{\text{rolling}} - 0.5|, \tag{6}$$

where $\Delta\alpha_{\text{surrogate}}$ quantifies true multifractality due to long-term correlations, any widening of this spectrum would imply inefficiency. In an efficient market, $H = 0.5$; any deviation indicates temporal correlations, with $H > 0.5$ signaling persistence and $H < 0.5$ indicating anti-persistence. This approach combines the classical Hurst exponent (order 2) and the multifractal spectrum (all orders) for a comprehensive measure of inefficiency.

See Figure 9 for the inefficiency index I computed on indexes.

5 Trading Strategy

To emphasize the practical implications of our inefficiency index, we propose a simple trading strategy based on a rolling hurst and the index's values. The strategy is based on the ssec since given the inefficiency plot it seems like the ssec might be the most inefficient series with highest inefficiency corresponding to crises of 2008 and 2015. If we relied solely on a simple momentum or mean reversion strategy based on the Hurst exponent, we would encounter numerous false signals and often hold positions for only one or two days. This would result in high transaction costs and poor performance. The purpose of our inefficiency index is to act as a filter for the Hurst signal, allowing us to take positions only when the market is broadly inefficient across all scales. The strategy is as follows:

1. Calculate the rolling Hurst exponent H_{rolling} using a window size of 6 months via the Modified R/S method.
2. Calculate the inefficiency index I using the formula defined in Section 4.1.
3. Set a threshold for the inefficiency index, denoted as $I_{\text{threshold}}$. This threshold is set on 1.5 standard deviation based on a rolling 6 months.
4. If $I > I_{\text{threshold}}$ and Hurst < 0.5, it indicates a potential market inefficiency, and mean reversion so we take a short position in the asset, otherwise we are long.

The performance can be found in Table 4, the graphical representation respectively in Figure 11 and Figure 12. The number of position taken is reduced with our index (421 vs 443), which showcases that our inefficiency index serves as a filter for the Hurst signal, reducing the number of false signals and drastically improving the overall performance of the strategy in terms of Sharpe ratio, Annualized return and Max Drawdown.

6 Conclusion

In this paper, we examined the long-term memory and multifractal properties of major stock market indices using both traditional and modified R/S analysis alongside MF-DFA. Our findings suggest that, while most series display Hurst exponents greater than 0.5 implying some degree of persistence the modified R/S approach indicates that only the Russell 2000 exhibits statistically significant long memory. We introduce an inefficiency index that gauges how far a market departs from efficiency by combining two elements: the width of its multifractal spectrum, which captures long-range correlations revealed through surrogate analysis, and the deviation of a rolling Hurst exponent from the benchmark value of 0.5, signalling momentum or mean-reversion effects. By integrating these components, our index simultaneously quantifies directional inefficiencies and multifractal complexity across all orders. This index can serve as a filter of the Hurst signal to reduce the number of false signals and improving performance.

7 Appendix

7.1 Demonstration of the covariance of fractional Brownian motion (fBm)

The fractional Brownian motion (fBm), denoted by $X_H(t)$, is defined as a zero-mean continuous-time Gaussian process whose increments are correlated. Its covariance function is given by:

$$C_H(t,s) = \frac{\sigma^2}{2}\left(t^{2H} + s^{2H} - |t-s|^{2H}\right)$$

where $H \in (0,1)$ is the Hurst exponent.

A fractional Brownian motion $X_H(t)$ with $X_H(0) = 0$ has increments that are normally distributed with zero mean, specifically:

$$X_H(t) - X_H(s) \sim \mathcal{N}(0, \sigma^2|t-s|^{2H})$$

Given that the process is centered (zero mean), the covariance is defined as:

$$C_H(t,s) = \mathrm{Cov}(X_H(t), X_H(s)) = \mathbb{E}[X_H(t)X_H(s)]$$

Using the following algebraic identity:

$$X_H(t)X_H(s) = \frac{1}{2}\left[X_H(t)^2 + X_H(s)^2 - (X_H(t) - X_H(s))^2\right]$$

the covariance becomes:

$$C_H(t,s) = \frac{1}{2}\left(\mathbb{E}[X_H(t)^2] + \mathbb{E}[X_H(s)^2] - \mathbb{E}[(X_H(t) - X_H(s))^2]\right)$$

We have by definition of fBm:

$$\mathbb{E}[X_H(t)^2] = \sigma^2 t^{2H}, \quad \mathbb{E}[X_H(s)^2] = \sigma^2 s^{2H}, \quad \mathbb{E}[(X_H(t) - X_H(s))^2] = \sigma^2|t-s|^{2H}$$

Substituting these into our covariance expression, we get:

$$C_H(t,s) = \frac{1}{2}\left(\sigma^2 t^{2H} + \sigma^2 s^{2H} - \sigma^2|t-s|^{2H}\right)$$

Factoring out the term σ^2, we arrive at the final covariance formula:

$$\boxed{C_H(t,s) = \frac{\sigma^2}{2}\left(t^{2H} + s^{2H} - |t-s|^{2H}\right)}$$

This covariance function entirely characterizes the dependence structure of fractional Brownian motion, revealing long-term correlation when $H > 0.5$ (persistence) and anti-correlation when $H < 0.5$ (anti-persistence). To comeback where you left off, see Section 3.

7.2 Augmented Dickey-Fuller Test

Ticker	P-Value on log prices	P-Value on log differentiated return
S&P 500	0.863	0.000
Russell 2000	0.695	0.000
FTSE 100	0.226	0.000
Nikkei 225	0.660	0.000
DAX	0.663	0.000

Table 1 – P-values from the Augmented Dickey-Fuller (ADF) test for stationarity. The P-value of log prices refers to the Augmented Dickey Fuller test (ADF) on log prices, while the P-value of log-differentiated prices indicates the ADF test on log-differentiated returns. The null hypothesis is non-stationarity. To come back where you left off, see Section 3.1

7.3 Definition (Time Domain)

A stationary process X_t is said to exhibit long-range dependence (long memory) if there exist constants

$$a \in (0,1), \quad c > 0,$$

such that its autocorrelation function $\rho(k)$ satisfies

$$\lim_{k \to \infty} \frac{\rho(k)}{c\, k^{-\alpha}} = 1 \tag{7}$$

where $\rho(k)$ is the autocovariance function, c is a constant (V.Mignon 2003). To comeback where you left off, see Section 3.2.

7.4 Simulation of Fractional Brownian Motion

In this simulation, we aim to generate fractional Brownian motion (fBm) to better understand how the autocorrelation decays as a function of the Hurst exponent H. By simulating paths for different values of H, we can observe how the memory and persistence properties of the process vary. To generate the fractional Brownian motion (fBm), we use a Cholesky decomposition-based approach. The covariance matrix of fBm is given by (3):

where H is the Hurst exponent, which determines the degree of long-term dependence in the process.

The steps of the simulation are as follows:

1. Define a time grid of N points between 0 and T.
2. Compute the covariance matrix using (3).
3. Apply Cholesky decomposition to obtain a lower triangular matrix L.
4. Generate a vector W of standard normal random variables.
5. Obtain the fBm path by computing $X = LW$.

The params used for this simulation are N = 1000 number of points, T = 1 day, Hurst exponents $H = 0.2, 0.35, 0.5, 0.65, 0.8$, the number of lag for the autocorrelation is 40.

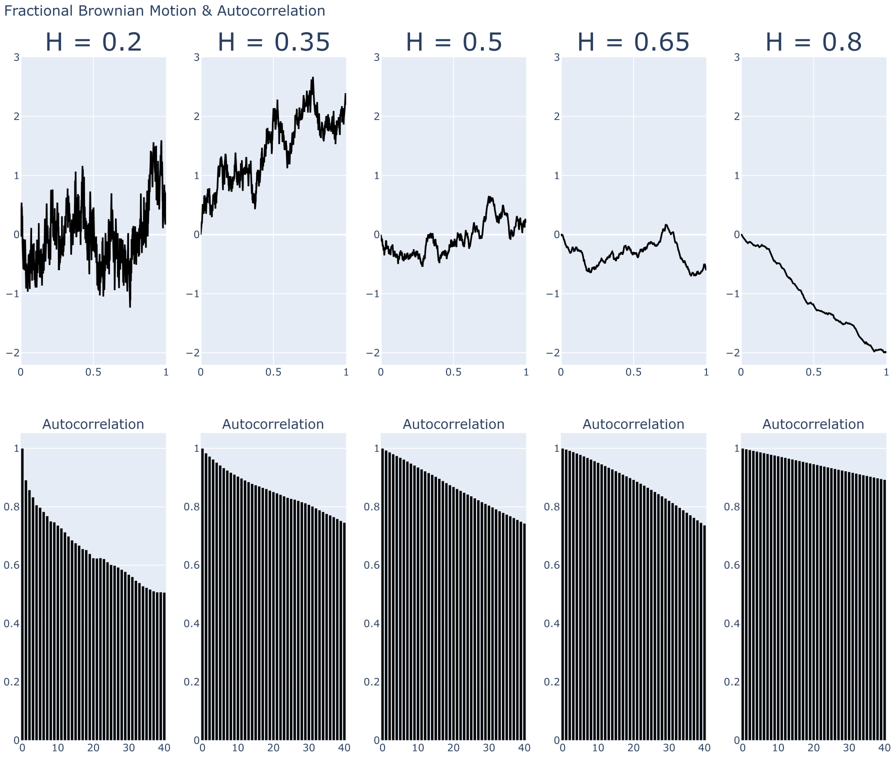

Figure 2 – Simulation of fractional Brownian motion with different Hurst exponent and its autocorrelation function. Value of H = 0.2, 0.35, 0.5, 0.65, 0.8 and 1000 points is simulated over a day.

The behavior of the fractional Brownian motion varies significantly with the Hurst exponent H.

When H is small (close to 0), the fBm exhibits high local variability, resulting in a highly granular trajectory with frequent fluctuations. The autocorrelation of increments decays rapidly, indicating that future values are weakly influenced by past values. This suggests a short-memory process, similar to standard Brownian motion.

As H increases, the autocorrelation decays more slowly, meaning that past values have a more significant impact on future values. This introduces a form of long-term dependence, where the process exhibits persistent trends. Consequently, the fBm trajectory appears smoother, with larger coherent movements and fewer abrupt changes.

In summary, a lower H leads to a more irregular and noisy path, characteristic of short-memory processes, while a higher H results in a smoother trajectory with stronger persistence.

7.5 R/S and Modified R/S Analysis

The R/S (Rescaled Range) analysis, introduced by Hurst and developed in various works by Mandelbrot, is certainly the most well-known method for estimating the Hurst exponent H. This statistic is defined as the range of the partial sums of deviations from the mean of a time series divided by its standard deviation. Consider a time series Y_t, $t = 1, \ldots, T$, with mean $\bar{Y}$. The range R is defined as:

$$R = \max_{1 \leq j \leq T} \left(Y_j - \bar{Y} \right) - \min_{1 \leq j \leq T} \left(Y_j - \bar{Y} \right). \tag{8}$$

The R/S statistic is then computed by dividing the range by the standard deviation s_T of the series:

$$Q_T = \frac{R}{s_T} = \frac{\max_{1\leq j\leq T}\left(Y_j - \bar{Y}\right) - \min_{1\leq j\leq T}\left(Y_j - \bar{Y}\right)}{s_T}, \quad \text{with } s_T = \sqrt{\frac{1}{T}\sum_{j=1}^{T}\left(Y_j - \bar{Y}\right)^2}. \tag{9}$$

Empirical studies by Mandelbrot and Wallis (1969b) have shown that Q_T scales with the number of observations T according to

$$Q_T \sim T^H, \tag{10}$$

which implies that by taking logarithms, the Hurst exponent H can be obtained from

$$H \sim \frac{\log(Q_T)}{\log(T)}. \tag{11}$$

Unfortunately, the asymptotic distribution of the R/S statistic is not known, making it difficult to establish a statistical test for the null hypothesis of short memory against the alternative hypothesis of long memory. Moreover, the R/S statistic does not explicitly account for short-term autocorrelation in the data, which can inflate (or reduce) the overall range and misrepresent the true variability of the series. Standard deviation estimates likewise ignore autocorrelated structure over short horizons, compounding the bias. As a result, the R/S measure can erroneously detect long memory when, in fact, short-term effects are responsible. This shortfall motivated Lo's Modified R/S procedure.

7.6 Modified R/S Analysis

The modified R/S statistic, denoted by $\tilde{Q}_T$, is defined as:

$$\tilde{Q}_T = \frac{R}{\hat{\sigma}_T(q)}, \tag{12}$$

where

$$\hat{\sigma}_T(q) = \sqrt{\frac{1}{T}\sum_{j=1}^{T}(Y_j - \bar{Y})^2 + \frac{2}{T}\sum_{j=1}^{T} w_j(q)\left[\sum_{i=j+1}^{T}(Y_i - \bar{Y})(Y_{i-j} - \bar{Y})\right]}, \tag{13}$$

and

$$w_j(q) = 1 - \frac{j}{q+1}. \tag{14}$$

This statistic differs from the traditional R/S statistic only by its denominator. In the presence of autocorrelation, the denominator does not solely represent the sum of the variances of the individual terms, but also includes autocovariances weighted according to lags q, with the weights $w_j(q)$ suggested by Newey and West (1987). Moreover, Andrews (1991) proposed a rule for choosing q:

$$q = [k_T] \quad \text{where} \quad k_T = \left(\frac{3T}{2}\right)^{\frac{1}{3}}\left(\frac{2\rho_1}{1-\rho_1^2}\right)^{\frac{2}{3}}, \tag{15}$$

where $[k_T]$ is the integer part of k_T, and ρ_1 is the first-order autocorrelation coefficient.

Unlike the classical R/S analysis, the limiting distribution of the modified R/S statistic is known. The statistic V, defined by

$$V = \frac{\tilde{Q}_T}{\sqrt{T}}, \tag{16}$$

converges to the range of a Brownian bridge over the unit interval. This convergence allows one to perform a statistical test for the null hypothesis of short memory against the alternative hypothesis of long memory by referring to the critical value table provided by Lo (1991), shown in Table 2. Therefore, accepting the null hypothesis implies that the series lacks the slow-decaying dependencies characteristic of long memory processes.

7.7 Critical Values for the Modified R/S Test

The critical values for the modified R/S test are provided in the table below. These values are used to assess whether the series exhibits long memory behavior based on the modified R/S statistic.

Significance Level	**critical value (modified R/S Statistic)**
0.005	2.098
0.05	1.747
0.10	1.620

Table 2 – Critical values for the modified R/S Statistic (Lo, 1991). To come back where you left off, see Section 3.2

The following table summarizes the results of the R/S statistic, modified R/S statistic, and the estimated Hurst exponents for each of the five indices analyzed:

Ticker	R/S	Hurst Exponent	Modified Hurst Exponent	Critical Value	Long Memory
S&P 500	30.166	0.558	0.501	1.007	False
Russell 2000	51.373	0.645	0.588	1.714	True
FTSE 100	40.236	0.605	0.548	1.341	False
Nikkei 225	22.234	0.508	0.508	1.048	False
DAX	26.985	0.540	0.540	1.278	False

Table 3 – Results for R/S, Hurst exponent, modified Hurst exponent, critical value at 10%, and rejection of the null hypothesis of no long memory from 1987-09-10 to 2025-02-28. The Hurst exponent can be equal for the R/S and modified R/S methods in the case where the autocorrelation coefficients are less than zero (refer to Section 7.3), in this case we set q equal to 0 and therefore the R/S and modified R/S share the same formula. To come back where you left off, see Section 3.2

8 Multifractal Detrended Fluctuation Analysis

The Multifractal Detrended Fluctuation Analysis (MF-DFA) is a generalization of the standard Detrended Fluctuation Analysis approach designed to detect multifractality in time series (Kantelhardt et al., 2002). The procedure can be summarized in five steps, as described below:

1. **Profile construction.** Given a series $\{x_k\}_{k=1}^{N}$, we first compute its mean $\bar{x}$. Then, we build the profile

$$Z(i) = \sum_{k=1}^{i} (x_k - \bar{x}), \quad i = 1, 2, \ldots, N, \tag{17}$$

where we use $Z(i)$ instead of $Y(i)$ to avoid confusion with previous definitions. This cumulative sum helps capture the local fluctuations in the data.

2. **Division into segments.** We split the profile $Z(i)$ into $N_s \equiv \lfloor N/s \rfloor$ non-overlapping segments, each of length s. Since N may not be a multiple of s, we repeat this procedure starting from the opposite end, yielding a total of $2N_s$ segments.
3. **Detrending.** For each of the $2N_s$ segments, we fit a polynomial trend (often linear or quadratic) and subtract it from $Z(i)$ in that segment. Let $z_\nu(i)$ be the fitting polynomial in segment ν. We then define the local variance as

$$F^2\bigl(s,\nu\bigr) \;=\; \frac{1}{s}\sum_{i=1}^{s}\Bigl[Z\bigl((\nu-1)s+i\bigr) \;-\; z_\nu(i)\Bigr]^2. \tag{18}$$

   This detrending step removes possible polynomial trends in the data.
4. **Generalized fluctuation function.** For each scale s, we compute the qth-order fluctuation function,

$$F_q(s) \;=\; \left\{\frac{1}{2N_s}\sum_{\nu=1}^{2N_s}\Bigl[F^2\bigl(s,\nu\bigr)\Bigr]^{q/2}\right\}^{1/q}. \tag{19}$$

   Varying q allows us to emphasize large ($q > 0$) or small ($q < 0$) fluctuations.

   In the special case $q = 0$, the fluctuation function is defined by a logarithmic averaging (see proof in Appendix Section 8.1):

$$F_0(s) = \exp\left(\frac{1}{4N_s}\sum_{\nu=1}^{2N_s}\ln\Bigl[F^2\bigl(s,\nu\bigr)\Bigr]\right). \tag{20}$$

5. **Scaling behavior.** Finally, on double-logarithmic axes, we examine the dependence of $F_q(s)$ on s. If

$$F_q(s) \;\sim\; s^{h(q)}, \tag{21}$$

   then $h(q)$ is called the generalized Hurst exponent. In a multifractal series, $h(q)$ varies with q, indicating different scaling behaviors for large versus small fluctuations.

For monofractal series, $h(q)$ is approximately constant for all q. In contrast, for multifractal series, $h(q)$ strongly depends on q, revealing heterogeneity in the scaling of fluctuations. For a graphical representation of the steps used in the MF-DFA, refer to Figure 3.

8.1 Proof of $F_0(s)$ as $q \to 0$

Proof of $F_0(s)$ as $q \to 0$

$$F_q(s) = \Bigl[\tfrac{1}{2N_s}\sum_{v=1}^{2N_s}\bigl(F_v^2(s)\bigr)^{q/2}\Bigr]^{1/q} \implies \ln F_q(s) = \frac{1}{q}\ln S(q),$$

where

$$S(q) = \frac{1}{2N_s}\sum_{v=1}^{2N_s} e^{\frac{q}{2}\ln F_v^2(s)}.$$

As $q \to 0$, $\ln S(q) \to 0$ and we apply L'Hôpital:

$$\lim_{q\to 0}\ln F_q(s) = \lim_{q\to 0}\frac{\ln S(q)}{q} = \frac{S'(q)}{S(q)}\bigg|_{q=0} = \frac{1}{4N_s}\sum_{v=1}^{2N_s}\ln F_v^2(s).$$

Exponentiating:

$$F_0(s) = \exp\Big[\tfrac{1}{4N_s}\sum_{v=1}^{2N_s} \ln F_v^2(s)\Big] = \exp\Big[\tfrac{1}{2N_s}\sum_{v=1}^{2N_s} \ln F_v(s)\Big].$$

Thus $F_0(s)$ is the geometric mean of the segment fluctuations. To comeback where you left off, see Section 8.

8.2 MF-DFA Graphical Representation

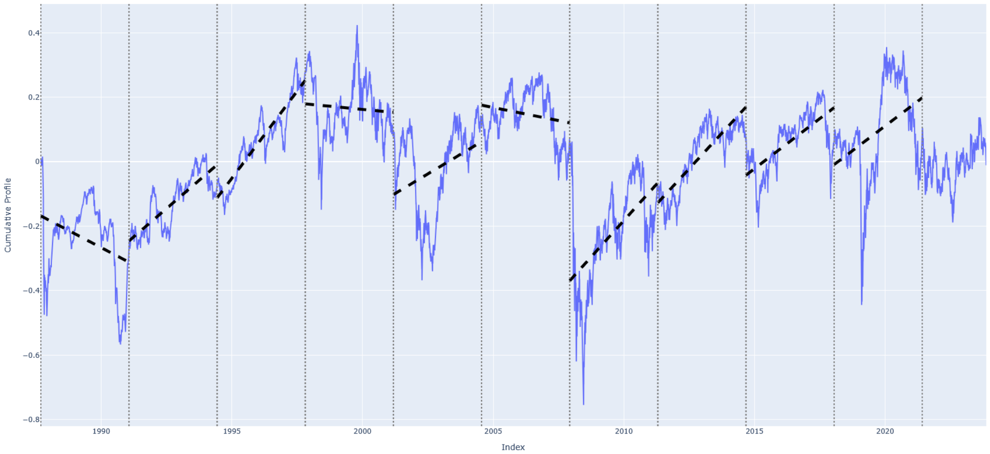

Figure 3 – Simulation of the MF-DFA steps, the series is the Russell 2000 returns split into segments of length 880. The blue line represents the cumulative sum of the centered returns, while the coloured lines represent the polynomial fit for each segment. To come back where you left off, see Section 8.

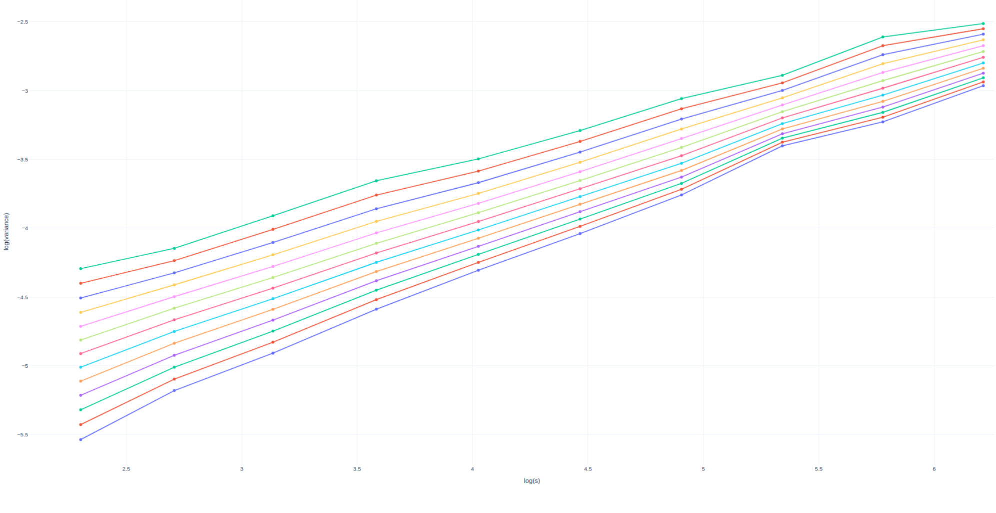

Figure 4 – Plot of the log scales (10 logly spaced increments from 10 to 500) against the log variance for each values of q, green line (highest line) represents q = -3 the lowest line represents q = 3. The slope of the line is the Hurst exponent for each q. The function is increasing with scale because, as the window size grows, larger fluctuations are aggregated, leading to higher variance.

8.3 Multifractal Spectrum of Russell Returns

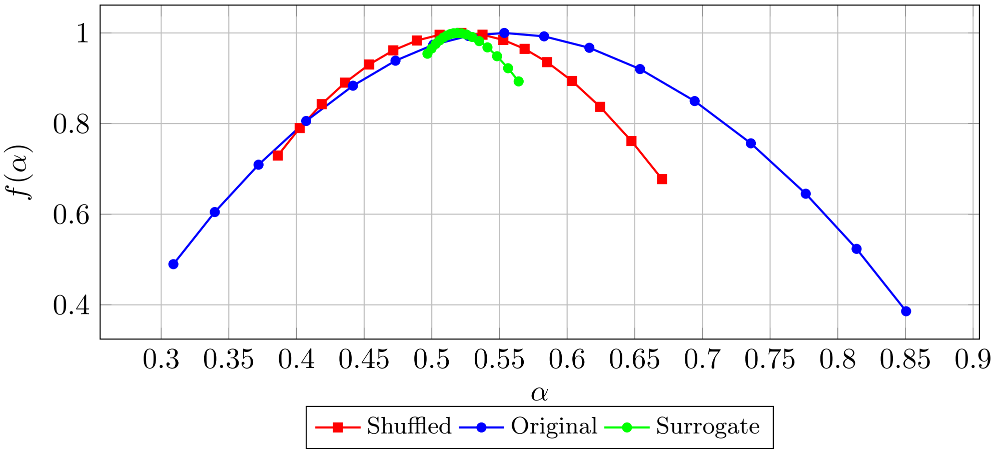

Figure 5 – Multifractal spectrum $f(\alpha)$ for the Russell 2000 returns. To comeback where you left off, see Section 3.5.

8.4 Multifractal Spectrum of S&P 500 Returns

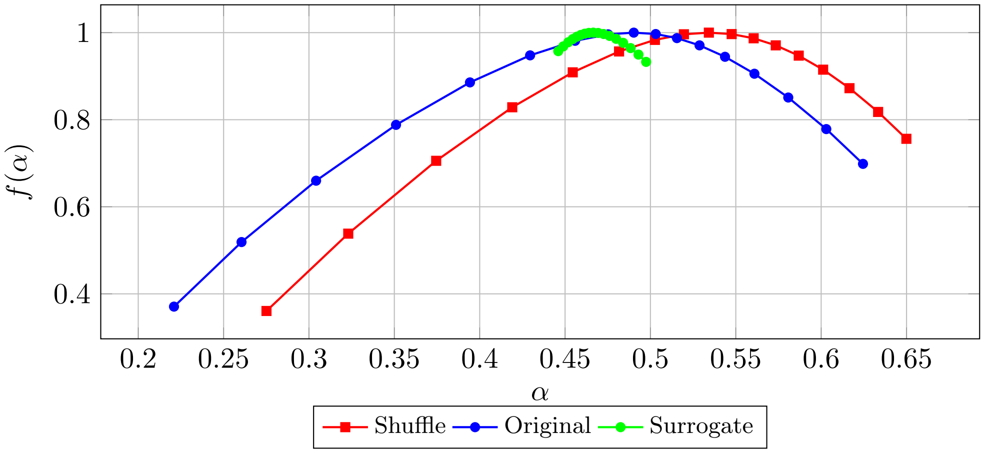

Figure 6 – Multifractal spectrum $f(\alpha)$ for S&P 500 returns.

The multifractal spectrum of the S&P 500 is noticeably narrower ($\Delta\alpha$= 0.40) and truncated on the right compared to the Russell 2000 ($\Delta\alpha = 0.55$), pointing to a weaker multifractal signature; its central peak at $\alpha = 0.486$ and an M–R/S Hurst exponent of 0.501 both underscore a near–random–walk dynamic. Unlike the Russell 2000, whose spectrum width shrinks when the series is shuffled—revealing the role of long-range correlations—the S&P 500's spectrum remains essentially unchanged by shuffling, indicating that its multifractality arises almost entirely from the non-Gaussian distribution of returns (kurtosis = 28.4). In practical terms, this means that extreme fluctuations in the S&P 500 cluster together but lack persistent temporal dependence, whereas the Russell 2000 exhibits genuine long-term memory. To come back where you left off, see Section 3.5.

9 Log Prices and Inefficiency Index

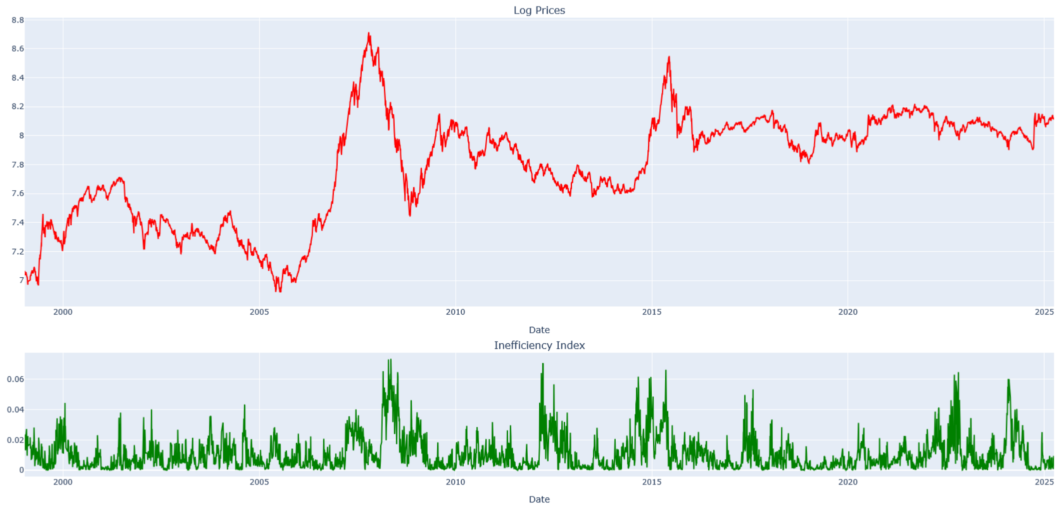

Figure 7 – Top : Log–price trajectory of the SSEC. Bottom : Inefficiency index

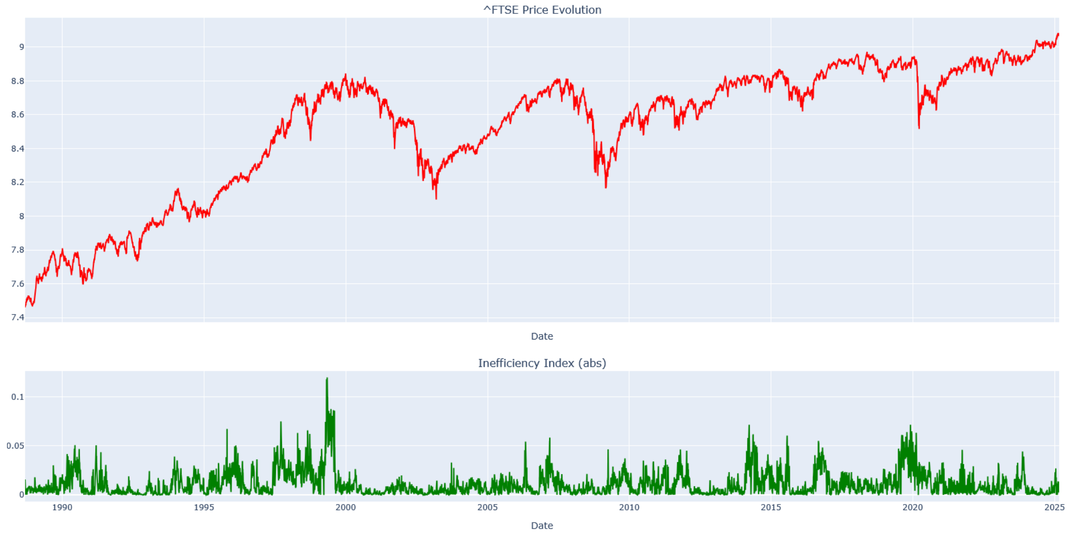

Figure 8 – Top : Log–price trajectory of the FTSE. Bottom : Inefficiency index

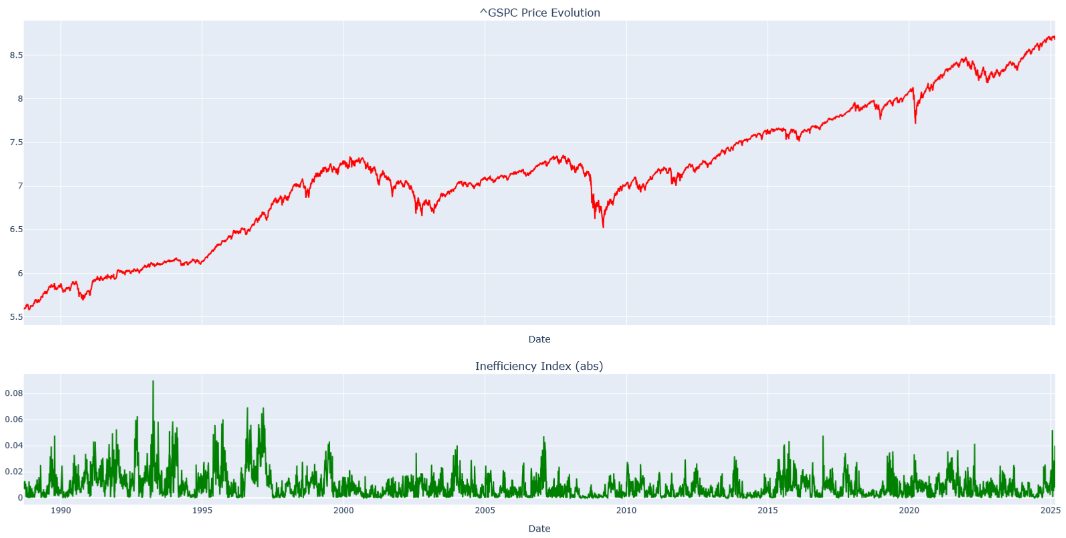

Figure 9 – Top : Log–price trajectory of the S&P 500. Bottom : Inefficiency index

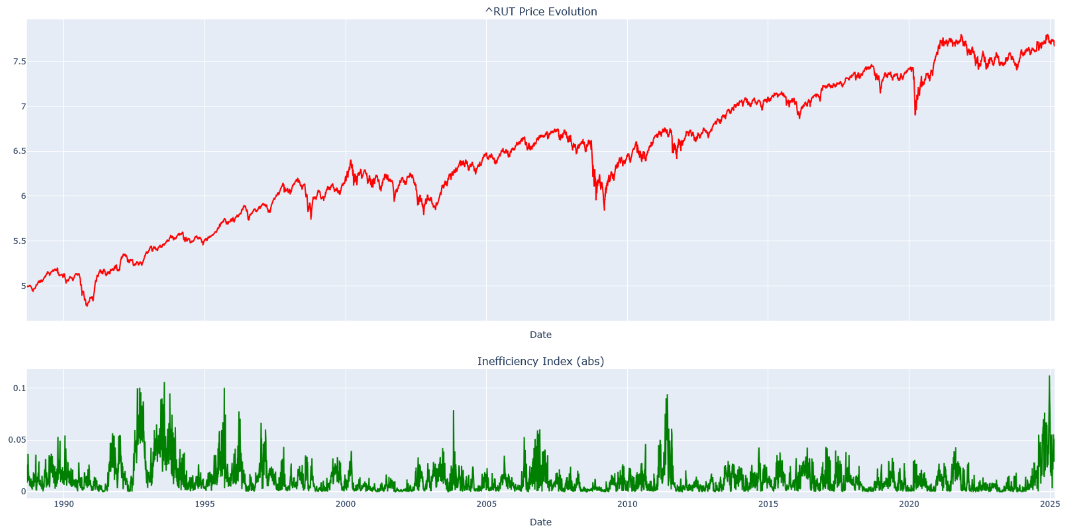

Figure 10 – Top : Log–price trajectory of the Russell 2000. Bottom : Inefficiency index

10 Trading Strategy

Strategy	Annualized Return	Annualized Volatility	Sharpe	Max Drawdown
Long/Short SSEC with inefficiency	9.521	23.307	0.409	-56.474
Long Only SSEC	3.326	23.316	0.143	-71.985
Long/short SSEC without inefficiency	5.075	23.314	0.218	-62.687

Table 4 – Performance metrics of the trading strategy with and without the inefficiency index on no transaction costs.

10.1 Cumulative Returns of the Trading Strategy

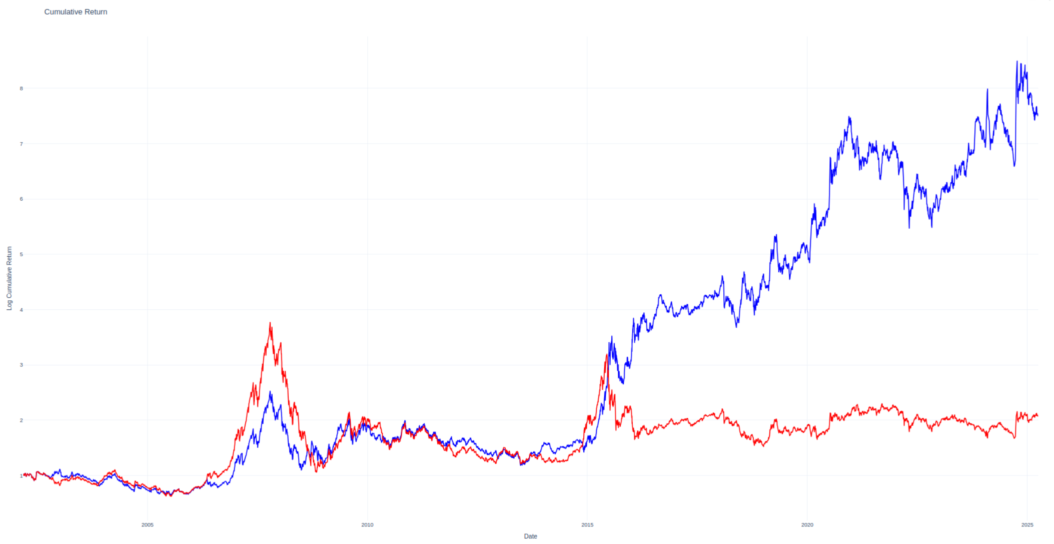

Figure 11 – Cumulative returns of the trading strategy with the inefficiency index. The blue line represents the cumulative returns of the strategy, while the red line represents the cumulative returns of the ssec. The strategy is based on the inefficiency index, which is used to filter Hurst signal.

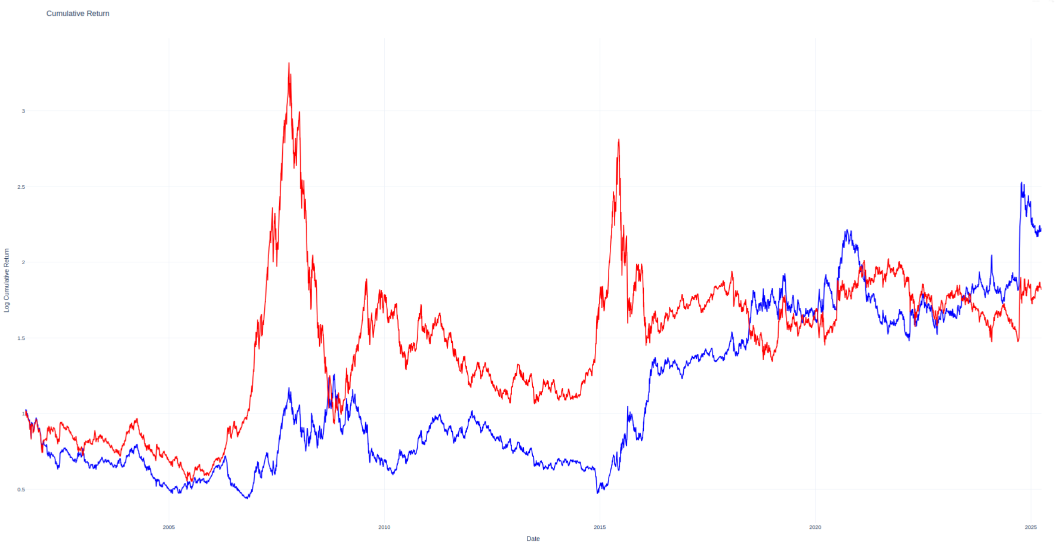

Figure 12 – Cumulative returns of the trading strategy without the inefficiency index. The blue line represents the cumulative returns of the strategy, while the red line represents the cumulative returns of the ssec. The strategy is long/short depending on the value of the Hurst (long > 0.5, short < 0.5), no filter are used.

11 References

Lo, A.W. (1991). *Long-Term Memory in Stock Market Prices*.

Mignon, V. (2003). *Méthodes d'estimation de l'exposant de Hurst. Application aux rentabilités boursières*, Économie & Prévision.

Kantelhardt, J.W., Zschiegner, S.A., Koscielny-Bunde, E., Bunde, A., Havlin, S., & Stanley, H.E. (2002). *Multifractal Detrended Fluctuation Analysis of Nonstationary Time Series. Physica A: Statistical Mechanics and its Applications*, 316(1–4), 87–114.

Lukasz Czarnecki, Dariusz Grech. (2009). *Multifractal dynamics of stock markets*. Acta Physica Polonica Series.

Kwapień, J., Drożdż, S., & Guhr, T. (2023). *Genuine multifractality in time series is due to temporal correlations: the effect of shuffling on multifractality*.

Andrews, D.W.K. (1991). *Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation*. *Econometrica*, 59(3), 817–858.

Mandelbrot, B.B. and Wallis, J.R. (1968). "Noah, Joseph, and Operational Hydrology", *Water Resources Research*, vol. 4, pp. 909–918.

Mandelbrot, B.B. (1973). "Le problème de la réalité des cycles lents et le syndrome de Joseph", *Economie Appliquée*, vol. 26, pp. 349–365.

Mandelbrot, B.B. and Wallis, J.R. (1969a). "Some Long-Run Properties of Geophysical Records", *Water Resources Research*, vol. 5, pp. 321–340.

Mandelbrot, B.B. and Wallis, J.R. (1969b). "Robustness of the Rescaled Range R/S in the Measurement of Noncyclic Long-Run Statistical Dependence", *Water Resources Research*, vol. 5, pp. 967–988.

Mandelbrot, B.B. and Taqqu, M.S. (1979). "Robust R/S Analysis of Long-Run Serial Correlation", *Bulletin of the International Statistical Institute*, vol. 48, pp. 69–104.

Di Matteo, T. (2007). *Multi-scaling in Finance*.