



UNIVERSITÀ
DI PAVIA

Dipartimento di Scienze Economiche e Aziendali
Corso di Laurea Magistrale in Finance

**FROM ECONOMETRIC FORECASTS
TO DYNAMIC PORTFOLIO
ALLOCATION: EVIDENCE FROM US
AND EUROPEAN SECTOR ETFS**

Relatore:

Chiar.mo Prof. Giacomo Bormetti

Tesi di Laurea
di Luca Bertelli

Matr. n. 557412

Anno Accademico 2025–2026

ABSTRACT

This thesis investigates whether parsimonious econometric forecasts contain economically useful information for sector ETF allocation. Weekly log returns for US and European sector ETF universes are modeled with autoregressive, moving-average, and ARMA specifications selected by the Bayesian Information Criterion. The forecasts are converted into transaction-cost-adjusted 130/30 portfolios. To prevent look-ahead bias, all model and strategy choices are made using a validation sample ending on December 31, 2019. For each region, the validation exercise fixes one static alpha, one time-varying alpha rule, and one mean-variance specification. These choices are then held unchanged in a true out-of-sample comparison beginning on January 3, 2020. In the US test period, the selected static-alpha portfolio achieves an annualized return of 24.09 percent and a Sharpe ratio of 0.958, compared with 15.64 percent and 0.807 for the market benchmark. In Europe, the selected static-alpha portfolio raises annualized return from 8.92 to 10.73 percent, but its Sharpe ratio of 0.498 is essentially equal to the market's 0.499. The evidence is therefore strongest for the US market and illustrates why validation-based selection must be separated from final performance assessment.

Table of Contents

Abstract	ii
List of Tables	vi
List of Figures	vii
1 Introduction, Literature Review and Research Positioning	1
1.1 Return predictability and out-of-sample instability	2
1.2 Statistical accuracy and economic value of forecasts	4
1.3 Cross-sectional return prediction and sector rotation	5
1.4 Forecast-informed portfolio allocation	6
1.5 Model selection, data snooping, and true out-of-sample evaluation	7
1.6 Research gap and contribution	9
2 Theoretical Framework: Econometric and Portfolio Foundations	11
2.1 Recursive estimation frameworks	11
2.2 Econometric forecasting models	13
2.2.1 Autoregressive models	13
2.2.2 Moving-average models	14
2.2.3 ARMA models	14
2.3 Model selection through the Bayesian Information Criterion	14

2.4	Portfolio theory and constrained allocation	15
2.4.1	Mean-variance portfolio theory	15
2.4.2	Short selling and portfolio constraints	15
2.4.3	The 130/30 portfolio structure	16
2.5	Linking econometric forecasts to portfolio allocation	17
3	Data	18
4	Methodology	21
4.1	Empirical design and benchmark framework	21
4.2	Econometric forecasting framework	23
4.3	Forecast-based portfolio construction	25
4.4	Alpha-selection rules	26
4.5	Benchmark construction	27
4.5.1	Passive market benchmarks	27
4.5.2	Mean-variance benchmark	28
4.5.3	Scarecrow benchmark	28
4.6	Transaction costs and evaluation metrics	29
5	Empirical Results	31
5.1	Forecasting diagnostics	31
5.2	Validation-period portfolio analysis	38
5.2.1	Portfolio weights and static alpha	38

5.2.2	Rolling versus expanding alpha rules	41
5.2.3	Mean-variance validation	43
5.2.4	Models frozen at the validation cutoff	44
5.3	Transaction costs in validation	44
5.4	True out-of-sample strategy comparison	46
5.4.1	United States	46
5.4.2	Europe	48
6	Discussion	50
7	Conclusion and Future Research	53
	Appendix A: ACF Diagnostics	55
	Appendix B: Portfolio Weight Evolution	57
	Bibliography	59

List of Tables

3.1	ETF universe and sector mapping	19
5.1	Average forecast diagnostics in the validation period	32
5.2	Average weekly cross-sectional signal slopes	35
5.3	Conditional weekly returns by volatility-adjusted signal tercile	36
5.4	Static-alpha selection using validation data only	41
5.5	Alpha-rule selection in the validation period	43
5.6	Mean-variance selection in the validation period	43
5.7	Specifications frozen on December 31, 2019	44
5.8	Final US true out-of-sample results	47
5.9	Final European true out-of-sample results	49

List of Figures

2.1	Schematic comparison of rolling and expanding estimation windows. Rolling estimation keeps a fixed-length window, while expanding estimation retains earlier observations and adds new ones.	12
4.1	ACF diagnostics for the matched US and European technology ETFs, XLK and EXV3.DE, shown separately by region.	24
5.1	Rolling and expanding forecasts versus actual weekly log returns for the matched technology ETFs during validation.	33
5.2	US realized returns conditional on the volatility-adjusted forecast-signal tercile. Error bars are 90 percent Newey–West confidence intervals.	36
5.3	European realized returns conditional on the volatility-adjusted forecast-signal tercile. Error bars are 90 percent Newey–West confidence intervals.	37
5.4	US rolling 130/30 portfolio weights for $\alpha = 0.2$ and $\alpha = 0.4$.	39
5.5	Annualized return and Sharpe ratio across the static-alpha grid.	40
5.6	US rolling and expanding alpha-selection paths.	41
5.7	European rolling and expanding alpha-selection paths.	42
5.8	Implied annualized transaction-cost drag across the static-alpha grid.	45

5.9	US true OOS annualized return and Sharpe-ratio comparison.	46
5.10	US true OOS cumulative portfolio paths.	47
5.11	European true OOS annualized return and Sharpe-ratio comparison.	48
5.12	European true OOS cumulative portfolio paths.	48
A.1	Complete US weekly log-return ACF diagnostics.	55
A.2	Complete European weekly log-return ACF diagnostics.	56
B.1	US rolling 130/30 portfolio weights across the static alpha grid.	57
B.2	European rolling 130/30 portfolio weights across the static alpha grid.	58

Chapter 1

Introduction, Literature Review and Research Positioning

Forecasting expected returns is difficult because asset returns are noisy, weakly persistent, and highly regime-dependent. Nevertheless, even weak predictive signals may be economically meaningful if they are converted into portfolio weights in a disciplined way. The empirical analysis evaluates that idea using weekly sector ETF returns in the United States and Europe from 2010 to 2026.

The research question is whether econometric models contain enough predictive information to support economically meaningful portfolio construction after realistic transaction costs. The analysis asks whether autoregressive and ARMA-type forecasts can be transformed into performance-enhancing trading signals, and whether forecast-driven portfolios improve on a classical 130/30 mean-variance allocation after costs.

This study makes an applied empirical contribution rather than proposing a new forecasting estimator. First, it evaluates whether BIC-selected low-order AR, MA, and ARMA models contain economically useful information despite weak point-forecast accuracy. Second, it introduces diagnostics aligned with the

portfolio mechanism by testing whether volatility-adjusted forecasts monotonically rank subsequent ETF returns. Third, it compares the transmission of these signals into constrained 130/30 portfolios across matched US and European sector universes. Finally, it separates specification selection from final evaluation through a validation-frozen true out-of-sample design. Static alpha, the rolling-versus-expanding alpha rule, and the rolling-versus-expanding mean-variance benchmark are selected without accessing the final test sample; the retained strategies are then compared with the S&P 500 and STOXX Europe 600 and with a naive scarecrow model.

The remainder of this chapter positions the empirical analysis within five related strands of research: the instability of return predictability, the distinction between statistical accuracy and economic value, cross-sectional sector prediction, forecast-informed portfolio allocation, and the control of data snooping through genuine out-of-sample evaluation. The discussion progressively narrows from the general difficulty of return forecasting to the specific research gap addressed by the thesis. The econometric and portfolio-theory foundations are presented separately in Chapter 2.

1.1 Return predictability and out-of-sample instability

Forecasting expected equity returns is difficult because realized returns are dominated by news and contain only a small conditional-mean component. The aggregate

equity-premium literature provides particularly strong evidence of this difficulty. Welch and Goyal [10] reassess a broad collection of predictors and find that most perform poorly and unstably out of sample relative to the prevailing historical mean. The extended evaluation of Goyal, Welch, and Zafirov [11], which examines both the original predictors and variables proposed in subsequent research through 2021, reaches a similarly cautious conclusion: many predictors lose significance or predictive accuracy when their samples are extended.

This evidence does not imply that all return forecasting is uninformative. Campbell and Thompson [12] show that economically motivated restrictions can allow small predictive improvements to generate utility gains, while forecast combinations can reduce model instability by pooling information across predictors [13]. The broader lesson is that apparent predictability must be assessed in real time and against simple benchmarks. In the present study, this motivates low-order specifications, rolling and expanding estimation windows, and a final evaluation period that is not used for strategy selection.

Most of this literature forecasts the aggregate market premium and studies market timing between equities and a risk-free asset. The present research instead asks whether the own return history of individual sector ETFs contains information about their relative subsequent performance. Aggregate predictability remains relevant as a warning about low signal-to-noise ratios and parameter instability, but the decision problem is cross-sectional rather than a binary decision about total

market exposure.

1.2 Statistical accuracy and economic value of forecasts

Forecast evaluation can be statistical or decision-based. Statistical measures such as RMSE assess the numerical distance between a forecast and its realization, while correlation and directional accuracy assess association and sign prediction. These measures are important, but they do not necessarily coincide with the objective of an investor. Granger and Pesaran [14] formalize the distinction between statistical and economic forecast evaluation by emphasizing that forecast value depends on the loss function and the decision for which the forecast is used.

The distinction is empirically important in return prediction. Cenesizoglu and Timmermann [15] show that models with relatively poor out-of-sample point accuracy can still add economic value when their forecasts guide portfolio decisions, and that the relation between statistical performance and economic value is generally weak. A small forecast difference can matter if it changes an economically consequential position, whereas a numerically accurate forecast can be irrelevant if it leaves the allocation unchanged.

This insight is central to the present research. RMSE and hit ratio evaluate each ETF forecast through time, but the 130/30 strategy uses the relative magnitudes of volatility-adjusted forecasts across ETFs. Its relevant intermediate question is therefore not only whether $\hat{r}_{i,t}$ is close to $r_{i,t}$, but whether assets with larger relative

signals subsequently realize larger returns. The empirical analysis connects three levels of evidence:

$$\begin{aligned} \text{point-forecast accuracy} &\longrightarrow \text{cross-sectional signal ordering} \\ &\longrightarrow \text{out-of-sample portfolio performance.} \end{aligned} \tag{1.1}$$

The cross-sectional slope and signal-tercile diagnostics in Chapter 5 are designed specifically to examine the middle link.

1.3 Cross-sectional return prediction and sector rotation

The cross-sectional literature provides a closer analogue to the portfolio mechanism used here than aggregate market-timing studies. Moskowitz and Grinblatt [16] document a strong industry component in momentum and show that strategies based on past winning and losing industries can contain economically relevant information. Beber, Brandt, and Kavajecz [17] find that institutional rebalancing across equity sectors is consistent with sector rotation and contains information about economic conditions. These studies establish that relative sector movements can be meaningful even when aggregate return forecasting remains difficult.

More recent work uses richer prediction methods to rank assets. Rapach, Strauss, Tu, and Zhou [18] use lagged returns across industries and machine-learning tools to produce economically valuable industry-rotation forecasts. Gu, Kelly, and Xiu [19] show more broadly that flexible prediction methods can generate economic gains in both time-series market timing and cross-sectional long-short portfolios, while

emphasizing the importance of genuine out-of-sample evaluation and controls against overfitting.

The study most directly related to the present application is Bessler and Wolff [20]. They forecast monthly US sector-index returns using macroeconomic, fundamental, and technical predictors and incorporate the forecasts into Black–Litterman portfolios. Their results also demonstrate that statistical forecast measures and portfolio value need not move together. The present analysis is complementary rather than a replication. It uses weekly investable ETFs, relies only on each asset’s return history, selects parsimonious AR, MA, and ARMA models with BIC, employs a direct volatility-scaled 130/30 mapping, and compares matched US and European sector universes. The objective is to determine whether a substantially simpler and more transparent information set can still produce a useful relative ordering of sector returns.

1.4 Forecast-informed portfolio allocation

The portfolio literature provides the decision framework through which return forecasts acquire economic meaning. Mean-variance theory treats allocation as a joint choice over expected return and risk [5, 6], with the efficient set and portfolio-separation properties developed further by Merton [7]. Forecasts can therefore be useful even when their numerical accuracy is modest, provided that they improve the relative allocation of capital across assets.

The practical difficulty is that expected returns and covariances are estimated with error. Small input errors can be amplified into unstable and concentrated weights, a problem emphasized by Michaud [8]. Jagannathan and Ma [9] show that constraints can reduce the effect of estimation error, while DeMiguel, Garlappi, and Uppal [21] demonstrate how difficult it is for estimated optimal portfolios to outperform naive diversification consistently out of sample. This literature motivates both the constrained mean-variance benchmark and the disciplined exposure limits adopted in the empirical analysis.

Lo and Patel [22] describe 130/30 portfolios as active-extension strategies that relax the long-only constraint and allow relative signals to be expressed on both the long and short sides. This flexibility does not create predictive content by itself: if the ranking signal is uninformative, greater gross exposure should not generate systematic value. A common 130/30 structure is therefore well suited to comparing forecast-based sector allocation with a constrained mean-variance benchmark. Chapter 2 formalizes these portfolio foundations and their connection with the econometric forecasts.

1.5 Model selection, data snooping, and true out-of-sample evaluation

Specification search creates a further challenge. When many predictors, models, or portfolio parameters are evaluated on the same observations, the apparent winner can reflect chance rather than persistent predictive content. White [23] formalizes this

data-snooping problem and shows why repeated reuse of a single historical sample can lead to overstated evidence of superior performance.

The risk is material in this application because the empirical procedure compares multiple ARMA orders, alpha values, information windows, and benchmark specifications. BIC penalizes model complexity within each forecasting window, but it does not by itself protect the final portfolio comparison from selection on realized strategy performance. The validation/test separation serves that distinct purpose. All choices among static alpha values, dynamic alpha rules, and mean-variance windows are completed using observations through December 31, 2019. Only the resulting frozen rules enter the independent period beginning January 3, 2020. Model parameters continue to be re-estimated using information available at each date, but no test-period realization is allowed to revise the selection protocol or the retained strategy families.

This design does not eliminate every form of specification uncertainty, because the researcher still observes only one historical path. It does, however, change the interpretation of the evidence: validation performance determines which rules would have been selected at the cutoff, while test performance measures how those predetermined rules subsequently behave.

1.6 Research gap and contribution

Existing research has extensively examined aggregate equity-premium prediction and, more recently, cross-sectional stock and industry return forecasting using large predictor sets and machine-learning methods. Less attention has been devoted to whether parsimonious forecasts based solely on an asset’s own return history can generate economically useful relative signals across investable sector ETFs. Moreover, statistical forecast evaluation and portfolio performance are often examined separately, even though constrained long-short allocation depends primarily on the cross-sectional ordering of predicted returns rather than on point-forecast accuracy.

This study addresses these issues by linking conventional forecast diagnostics, cross-sectional signal-ranking tests, and validation-frozen out-of-sample 130/30 portfolio performance across matched US and European sector ETF universes. Its contribution is applied and comparative rather than the proposal of a new estimator. First, it evaluates whether BIC-selected low-order AR, MA, and ARMA models contain economically useful information despite weak point-forecast accuracy. Second, it introduces diagnostics aligned with the allocation mechanism by testing whether volatility-adjusted forecasts monotonically rank subsequent ETF returns. Third, it studies whether the same transparent forecasting and portfolio framework produces comparable evidence across two regional sector universes. Fourth, it separates strategy selection from final performance assessment through a strict validation/test design.

Accordingly, the empirical claim is deliberately narrower than a general claim that ARMA models predict ETF returns. The research evaluates whether weak and parsimonious econometric forecasts contain economically useful cross-sectional information when embedded in a constrained sector-allocation framework, and whether any such value survives an independent test period.

Chapter 2

Theoretical Framework: Econometric and Portfolio Foundations

This chapter presents the established theoretical tools used in the empirical analysis. It first defines the recursive information frameworks under which forecasts are produced, then introduces the linear time-series models and the criterion used to select among them. The final sections review mean-variance theory and the 130/30 structure before explaining the conceptual link between return forecasts and constrained portfolio allocation.

2.1 Recursive estimation frameworks

Out-of-sample forecasting requires the information set to be defined separately at every forecast date. Let $\mathcal{I}_t$ denote the observations available when the forecast for period $t + 1$ is formed. A recursive procedure estimates the model using only $\mathcal{I}_t$, produces the next forecast, and updates the estimation sample only after the corresponding realization becomes available. This ordering prevents future observations from entering either parameter estimation or model selection.

Two recursive information structures are considered. An expanding window

retains the complete history available at each date. If the initial estimation sample contains observations $1, \dots, T_0$, the next update adds the newly observed return and estimates the model using $1, \dots, T_0 + 1$. A rolling window instead keeps a fixed number W of observations: at date t , estimation uses $t - W + 1, \dots, t$ and removes the oldest observation when a new one is added.

The distinction reflects a stability–adaptability trade-off. Expanding estimation uses progressively more observations, which can reduce parameter uncertainty when the data-generating process is stable. Rolling estimation gives greater influence to recent observations and can adapt more quickly when return dynamics change, at the cost of estimating parameters from a smaller sample. Neither framework is uniformly preferable, so the empirical analysis compares them under the same forecasting and evaluation protocol.

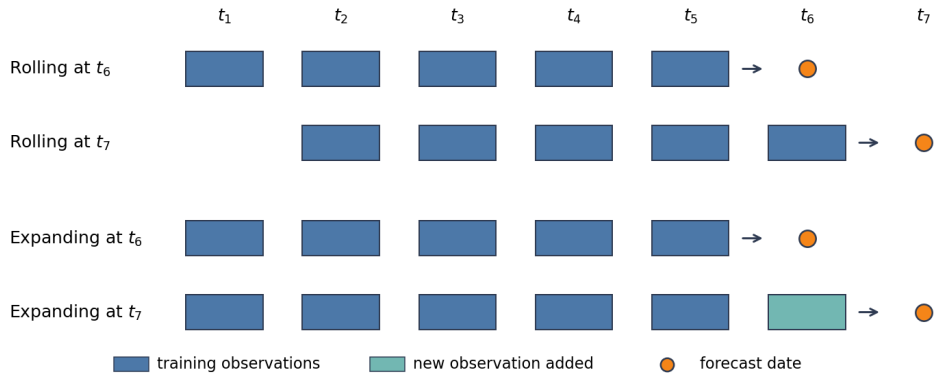


Figure 2.1: Schematic comparison of rolling and expanding estimation windows. Rolling estimation keeps a fixed-length window, while expanding estimation retains earlier observations and adds new ones.

2.2 Econometric forecasting models

Time-series models provide a systematic framework for describing how observations evolve and for assessing whether past information contains predictive structure. Returns are generally more suitable than price levels for this purpose because they are comparable across assets and closer to stationarity. The central object is the conditional distribution of the next return given the information available at the forecast date, with particular attention to its conditional mean [2, 3].

AR, MA, and ARMA models belong to the linear stochastic time-series framework developed in the Box–Jenkins tradition [1]. Under weak stationarity, the unconditional mean and variance are constant and autocovariances depend only on the lag. Financial returns usually contain limited linear dependence, however, so parsimonious model orders and repeated out-of-sample estimation are essential.

2.2.1 Autoregressive models

An autoregressive model of order p , $\text{AR}(p)$, represents current returns as a linear function of their lagged values:

$$r_t = c + \sum_{i=1}^p \phi_i r_{t-i} + \varepsilon_t, \quad (2.1)$$

where c is a constant, ϕ_i are autoregressive coefficients, and ε_t is a white-noise innovation. The model produces a predictive signal when lagged returns contain information about the next conditional mean.

2.2.2 Moving-average models

A moving-average model of order q , $\text{MA}(q)$, represents the return as a function of current and lagged innovations:

$$r_t = \mu + \varepsilon_t + \sum_{j=1}^q \theta_j \varepsilon_{t-j}, \quad (2.2)$$

where μ is the unconditional mean and θ_j measures the propagation of past shocks. This specification captures temporary innovation dynamics rather than direct persistence in observed returns.

2.2.3 ARMA models

An $\text{ARMA}(p, q)$ model combines both mechanisms:

$$r_t = c + \sum_{i=1}^p \phi_i r_{t-i} + \varepsilon_t + \sum_{j=1}^q \theta_j \varepsilon_{t-j}. \quad (2.3)$$

Low-order ARMA models can describe short-memory dynamics without the large parameter sets that would be especially vulnerable to overfitting in weekly ETF returns. AR and MA models are nested special cases of this broader family.

2.3 Model selection through the Bayesian Information Criterion

The Bayesian Information Criterion selects among competing time-series specifications:

$$\text{BIC} = -2\log(\hat{L}) + k \log(n), \quad (2.4)$$

where $\hat{L}$ is the maximized likelihood, k is the number of estimated parameters, and n is the sample size. Because its penalty increases with model complexity and sample size, BIC favors parsimonious specifications [4]. In the empirical implementation, it is recalculated for every ETF and re-estimation date, allowing the selected short-memory structure to differ across assets and over time.

2.4 Portfolio theory and constrained allocation

2.4.1 Mean-variance portfolio theory

Mean-variance theory formalizes portfolio choice as a joint decision about expected return and risk [5, 6]. With expected-return vector μ , covariance matrix Σ , and risk-aversion parameter λ , a standard quadratic objective is

$$\max_w \mu'w - \frac{\lambda}{2}w'\Sigma w. \quad (2.5)$$

For a given set of inputs, the efficient frontier contains portfolios with the lowest variance for a target expected return, or the highest expected return for a target variance. Its analytical properties and relation to portfolio separation were developed further by Merton [7].

2.4.2 Short selling and portfolio constraints

Estimated mean-variance portfolios can be highly sensitive to small errors in expected returns and covariances, producing unstable or concentrated allocations [8]. Constraints limit the feasible set and can therefore moderate the transmission

of estimation error into portfolio weights [9]. They also make comparisons more interpretable by ensuring that competing allocation rules operate under common exposure limits.

Allowing short positions expands the opportunity set because negative views can be expressed directly rather than only through underweighting. It also raises gross exposure and implementation risk, however, so the long and short books must be controlled explicitly. The theoretical role of the constraints is therefore not merely numerical: they determine how strongly a forecast or expected-return estimate can affect the final allocation.

2.4.3 The 130/30 portfolio structure

Lo and Patel [22] describe 130/30 portfolios as active-extension strategies that combine a 130 percent long book with a 30 percent short book. The structure retains net market exposure of 1.0 while increasing gross exposure to 1.6:

$$\mathbf{1}'w = 1.0, \tag{2.6}$$

$$\sum_{i:w_i>0} w_i = 1.3, \tag{2.7}$$

$$\sum_{i:w_i<0} w_i = -0.3, \tag{2.8}$$

$$\sum_i |w_i| = 1.6. \tag{2.9}$$

The additional long capacity is financed by the short book. This permits stronger expression of relative views without changing the portfolio's net exposure. Using

identical targets for the forecast-based and mean-variance portfolios makes it possible to distinguish the value of the econometric signal from the general effects of leverage and short selling.

2.5 Linking econometric forecasts to portfolio allocation

The two parts of the framework are connected through relative expected returns. At each date, the recursive estimation window defines the available information, BIC selects a parsimonious model for each ETF, and that model produces a one-step-ahead return forecast. Because ETFs have different risk levels, the forecasts are made comparable through volatility adjustment and interpreted cross-sectionally rather than as isolated point predictions. Assets with stronger relative signals receive larger allocations, while weaker signals are assigned smaller or negative exposures within the 130/30 constraints.

This sequence clarifies the role of each component. The estimation framework governs the balance between historical stability and adaptation, the econometric model extracts conditional-mean information, and the portfolio rule determines how that information is translated into economically meaningful positions. The next chapters describe the data and the exact empirical implementation of this mapping.

Chapter 3

Data

The data consist of weekly observations for US and European sector ETF universes, transformed into log returns. Each weekly price is the adjusted market close on Friday or, when no Friday observation is available because of a market holiday or closure, the last available closing price within that week. Let $P_{i,t}$ denote this weekly closing price for asset i . The corresponding continuously compounded return is

$$r_{i,t} = \log \left(\frac{P_{i,t}}{P_{i,t-1}} \right), \quad (3.1)$$

which converts proportional price changes into an additive return series suitable for time-series analysis [2]. The project uses 853 weekly return observations for each market, from 2010-01-15 to 2026-05-15. The benchmark series are SPY, which tracks the S&P 500, for the United States and EXSA.DE, the iShares STOXX Europe 600 UCITS ETF, for Europe [31].

Sector ETFs are used because they provide broad market exposure, high liquidity, and economically interpretable sector-level signals. The US universe is mainly built from Select Sector SPDR ETFs issued by State Street, with Vanguard ETFs used for Communication Services and Real Estate. The European universe is built from iShares STOXX Europe 600 sector ETFs issued by BlackRock [32]. These are large ETF families with exchange-traded shares, diversified underlying baskets,

and long histories for most sectors, which makes them suitable for testing whether weak time-series forecasts can be transformed into portfolio tilts while reducing idiosyncratic single-stock noise.

Table 3.1: ETF universe and sector mapping

Sector	US ticker	EU ticker
Benchmark	SPY	EXSA.DE
Technology	XLK	EXV3.DE
Health Care	XLV	EXV4.DE
Industrials	XLI	EXH4.DE
Financials / Banks	XLF	EXV1.DE
Energy	XLE	EXH1.DE
Consumer Discretionary	XLY	EXH8.DE
Consumer Staples	XLP	EXH7.DE
Materials	XLB	EXV6.DE
Utilities	XLU	EXH9.DE
Communication Services	VOX	EXV2.DE
Real Estate	VNQ	EXI5.DE

The two US exceptions are driven by data availability rather than by a different investment objective. Communication Services uses Vanguard Communication Services ETF (VOX) because the comparable SPDR fund XLC was launched only

on Jun 18, 2018 [27], while VOX has an inception date of Sep 23, 2004 [29]. Real Estate uses Vanguard Real Estate ETF (VNQ) because XLRE was launched on Oct 07, 2015 [28], while VNQ also dates back to Sep 23, 2004 [30]. Since the empirical sample begins in January 2010, VOX and VNQ provide continuous pre-sample history and avoid shortening the backtest.

Chapter 4

Methodology

4.1 Empirical design and benchmark framework

The empirical analysis evaluates whether one-step-ahead econometric forecasts can improve sector ETF allocation after accounting for risk scaling and transaction costs. All portfolio decisions are made using information available before the corresponding return is realized. Following the recursive frameworks defined in Section 2.1, the rolling implementation always uses the most recent 260 weekly observations, whereas the expanding implementation retains the complete history after the same initial five-year sample. The first forecasts therefore follow 260 training observations, and models and portfolio parameters are updated every 26 weeks, approximately every six months.

The empirical design then separates model development from final evaluation. The validation period contains all portfolio observations up to and including December 31, 2019. Every table and figure used to compare candidate specifications is truncated at this date. The test period begins with the first weekly observation strictly after the cutoff, January 3, 2020. No test-period return is used to choose an alpha, an alpha-selection window, or a mean-variance estimation window.

Selection is performed separately for the US and European universes. Within

each region, the validation Sharpe ratio determines one representative from each strategy family: one static-alpha forecast portfolio, one time-varying alpha rule (rolling or expanding), and one 130/30 mean-variance portfolio (rolling or expanding). Once selected, each specification is frozen. The strategy is run continuously so that weights accumulated before the cutoff carry into the test period, but performance statistics in the final comparison use only returns strictly after the cutoff. This design prevents the final sample from influencing either model choice or portfolio initialization.

The frozen forecast-driven portfolios are evaluated against three comparison categories in the final test period.

The market benchmarks are the S&P 500 for the United States and the STOXX Europe 600 for Europe. The S&P 500 contains 500 leading US large-cap companies and is weighted by float-adjusted market capitalization, covering approximately 80 percent of available US equity-market capitalization [25]. The STOXX Europe 600 contains 600 companies from 17 developed European countries, selected and weighted primarily by free-float market capitalization, and spans large-, mid-, and small-cap segments [26]. These indices establish whether active sector allocation adds value relative to broad long-only market exposure.

The 130/30 mean-variance portfolio is included to provide a valid comparison with another leveraged long-short strategy. Unlike the S&P 500 and STOXX Europe 600, which contain no short component, mean-variance operates under the same 130

percent long and 30 percent short exposure targets as the forecast-based strategy. It therefore helps separate the value of the ARMA signal from the more general effect of leverage and long-short portfolio construction.

The scarecrow is not a conventional market benchmark. It is a diagnostic model that replaces each ARMA forecast with the most recently observed return while keeping the remaining allocation procedure unchanged. Its role is to determine whether ARMA forecasts play a fundamental part in the strategy or whether similar results could be obtained from a naive signal combined with the same volatility scaling and 130/30 mapping. Detailed construction of all three comparisons is presented after the forecast-based portfolio and alpha-selection rules.

4.2 Econometric forecasting framework

The forecasting framework begins with autocorrelation-function diagnostics for each ETF. Figure 4.1 shows, for example, that the technology sector has a substantially different autocorrelation structure in the United States and Europe. Focusing on the first few lags, XLK and EXV3.DE display clearly different short-run dependence patterns. This suggests that the two markets, and more broadly the two regions, should not be expected to follow identical return dynamics. The complete ACF diagnostics for all ETFs are reported in Appendix A. The ACF plots are diagnostic rather than model-selection devices; final model selection is automated with BIC.

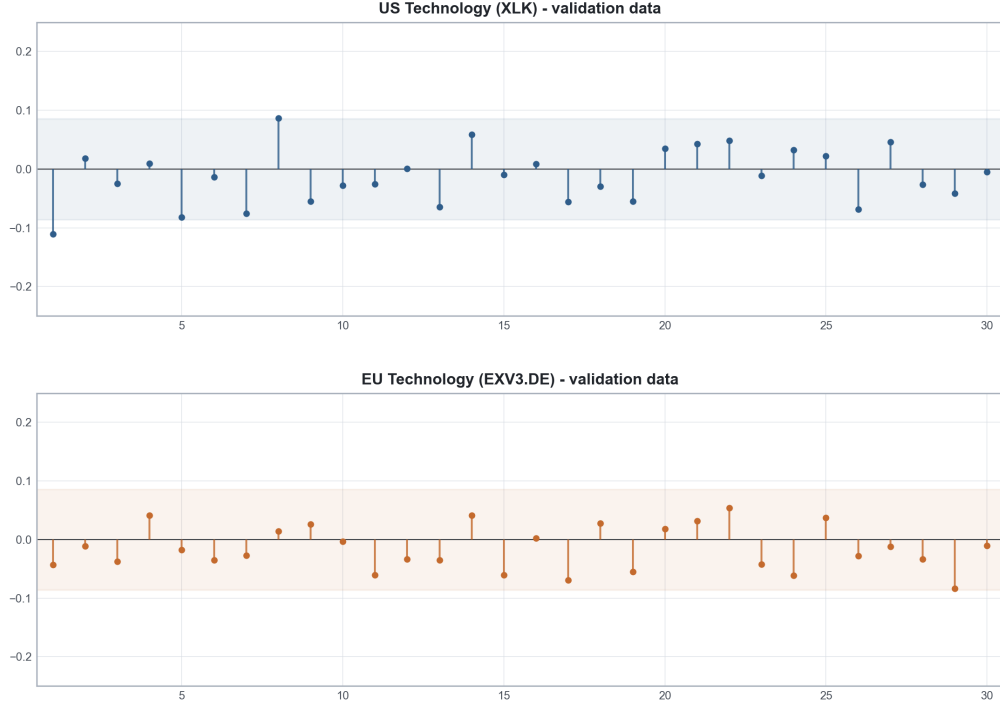


Figure 4.1: ACF diagnostics for the matched US and European technology ETFs, XLK and EXV3.DE, shown separately by region.

The candidate model set is:

$$\begin{aligned} \mathcal{M} = \{ & \text{AR}(1), \text{AR}(2), \text{AR}(4), \text{MA}(1), \text{MA}(2), \\ & \text{ARMA}(1, 1), \text{ARMA}(1, 2), \text{ARMA}(2, 1), \text{ARMA}(2, 2) \}. \end{aligned} \quad (4.1)$$

These models are deliberately low-order because the data are weekly and the objective is to avoid overfitting noisy return dynamics.

BIC selection has several practical advantages in this setting. It automates model choice without discretionary intervention, makes the choice data-driven rather than subjective, and makes the forecasting procedure reproducible because the same

candidate set and criterion are applied to every ETF and every re-estimation date. The main limitation is that a model is always selected from the candidate set, even when the true predictive signal is very weak or absent. For this reason, the selected model should be interpreted as the best available parsimonious specification within the grid, not as proof that the ETF return series is strongly predictable.

At each re-estimation date, BIC is recalculated separately for every ETF using the relevant rolling or expanding sample. The selected specification then produces one-step-ahead weekly return forecasts until the next scheduled re-estimation. This procedure allows model choice to vary across assets and over time while preserving a common, fully out-of-sample forecasting protocol.

4.3 Forecast-based portfolio construction

Forecasts are converted into portfolio weights with a forecast-magnitude rule. Realized volatility is estimated for each asset over the previous 52 weeks and shifted by one period, so that only information available before the portfolio date is used.

Let $\hat{\sigma}_{i,t}$ denote this lagged rolling volatility and let

$$s_{i,t} = \frac{\hat{r}_{i,t}}{\hat{\sigma}_{i,t}} - \frac{1}{N} \sum_{j=1}^N \frac{\hat{r}_{j,t}}{\hat{\sigma}_{j,t}}. \quad (4.2)$$

The raw portfolio update is then

$$\tilde{w}_{i,t} = w_{i,t-1} + \alpha s_{i,t}, \quad (4.3)$$

where $\hat{r}_{i,t}$ is the one-step-ahead return forecast and N is the number of ETFs in the regional universe. Dividing by realized volatility makes forecast magnitudes

comparable across assets with different risk levels. Cross-sectional centering makes the signal relative: assets with above-average signals receive larger raw scores, while assets below the average receive smaller raw scores. The implementation then maps the raw weights $\tilde{w}_{i,t}$ to a 130/30 allocation. Positive weights are rescaled to a total long exposure of 1.30 and negative weights to a total short exposure of -0.30, producing gross exposure 1.60 and net exposure 1.00.

At time zero, the portfolio is initialized as an equal-weighted allocation across the ETFs in each regional universe. The forecast rule therefore starts from a neutral sector allocation and subsequently introduces active tilts only as new one-step-ahead forecasts become available.

4.4 Alpha-selection rules

In the portfolio update defined by Equation 4.3, the parameter α controls signal aggressiveness. Low α values create modest changes in raw scores and allocations. High α values produce larger reallocations, higher turnover, and greater sensitivity to forecast noise. The static and dynamic rules use the grid

$$\{0.01, 0.05, 0.10, 0.15, 0.18, 0.20, 0.22, 0.25, 0.30, \\ 0.35, 0.40, 0.50, 0.70, 1.00, 1.50, 2.00, 3.00, 5.00\}.$$

The rolling alpha rule uses a 26-week internal selection window, re-selects every 26 weeks, and chooses the smallest alpha whose after-cost Sharpe is at least 95 percent of the best Sharpe in that internal window. The 95 percent threshold is intended

to reduce overfitting: when several alpha values deliver nearly indistinguishable performance, it favors the less aggressive and more parsimonious specification rather than selecting a larger alpha because of a small, potentially noisy Sharpe advantage.

The expanding alpha rule begins after the same minimum selection window, re-selects every 26 weeks, and chooses the alpha with the best expanding-sample after-cost Sharpe ratio, using the smaller alpha to break exact ties. Unlike the rolling rule in the current implementation, it does not apply the 95 percent near-optimality threshold.

These internal rules are distinct from the final validation-stage choice. At the cutoff, the complete validation returns of the rolling and expanding alpha rules are compared and only the rule with the higher validation Sharpe ratio is retained for the true out-of-sample test. Static alpha is also fixed at the cutoff. The static selector identifies the best validation Sharpe and then chooses the smallest alpha reaching 95 percent of that value, which imposes the same preference for a less aggressive near-optimal specification.

4.5 Benchmark construction

4.5.1 Passive market benchmarks

The passive benchmarks are SPY for the United States and EXSA.DE for Europe. Their returns are measured over the same evaluation period as the active strategies. These broad-market portfolios provide the baseline for determining whether the

additional complexity, turnover, and long-short exposure of the active allocation rules translate into economically meaningful improvements.

4.5.2 Mean-variance benchmark

The benchmark allocation is a 130/30 mean-variance portfolio. The expanding implementation begins with the initial 260-week sample and enlarges its estimation set over time. The rolling implementation uses the most recent 260 weekly observations. Both re-estimate every 26 weeks and solve the mean-variance problem with risk aversion $\lambda = 5.0$, gross exposure 1.6, net exposure 1.0, and a covariance ridge of 10^{-6} . The value $\lambda = 5.0$ is used as a moderate risk-aversion setting: it penalizes variance strongly enough to avoid extremely concentrated portfolios, while still allowing expected-return estimates to affect allocation. The ridge term is included only for numerical stability when covariance estimates are close to singular. Rolling and expanding mean-variance returns are compared only through the validation cutoff; the higher-Sharpe specification is then frozen for the final test.

4.5.3 Scarecrow benchmark

The scarecrow portfolio is a deliberately naive benchmark designed to test whether the econometric forecasts add information beyond a trivial extrapolation rule. For every ETF, its forecast is set equal to the most recently observed weekly return. This lagged-return signal is divided by the same 52-week lagged realized volatility and cross-sectionally centered in exactly the same way as the econometric signal. The

portfolio then applies the same 130/30 mapping and transaction-cost assumption. To isolate the contribution of the forecasting model, its alpha is matched to the region-specific value selected for the econometric static strategy: $\alpha = 0.22$ in the United States and $\alpha = 1.00$ in Europe. These values are fixed before the test period and are not re-optimized for the scarecrow. Consequently, the comparison holds signal aggressiveness constant within each region and attributes performance differences more directly to the forecast signal itself.

4.6 Transaction costs and evaluation metrics

Transaction costs are set to 5 basis points per unit of turnover, implemented as $0.0005 \times \text{turnover}$. This simple and transparent assumption is consistent with the view that sector ETFs are highly liquid instruments and that institutional execution can be meaningfully below retail bid-ask cost estimates. Transaction costs are included because forecast-driven strategies can generate economically relevant turnover, and ignoring implementation frictions would overstate realized performance. The 5 bps assumption is also in line with the broader empirical trading-cost literature emphasizing that costs must be incorporated directly into strategy evaluation [24].

Portfolio performance is evaluated using cumulative return, annualized return, annualized volatility, the Sharpe ratio, maximum drawdown, and turnover. Cumulative and annualized returns measure total and average yearly growth, while annualized volatility and maximum drawdown capture different dimensions of risk.

The Sharpe ratio summarizes return per unit of volatility, and turnover measures the amount of portfolio reallocation that generates transaction costs. Identical definitions are used in validation and test, but the two periods are never pooled for model selection or final reporting. Validation metrics determine the frozen specifications; final reported strategy performance is calculated solely from test-period observations.

Chapter 5

Empirical Results

This chapter maintains the separation between model development and final evaluation established in Chapter 4. Unless a result is explicitly identified as true out-of-sample, every diagnostic, sensitivity analysis, table, and figure used to evaluate or select a specification is calculated using observations no later than December 31, 2019. These validation results describe the evidence available when the portfolio rules are frozen; they are not part of the final test, which begins on January 3, 2020.

5.1 Forecasting diagnostics

The ACF diagnostics show that serial dependence is heterogeneous across ETFs. This supports the use of asset-specific model selection rather than imposing a common model on all sectors. The selected BIC models vary over time and across assets, indicating that return dynamics are unstable and regime-dependent.

Table 5.1 reports conventional point and directional forecast diagnostics. RMSE differences between rolling and expanding estimation are very small. Expanding estimation has the lower US RMSE, 0.0213 rather than 0.0214, whereas rolling estimation has the lower European RMSE, 0.0251 rather than 0.0252. Average forecast-return correlations are also close to zero. None of the positive correlations is

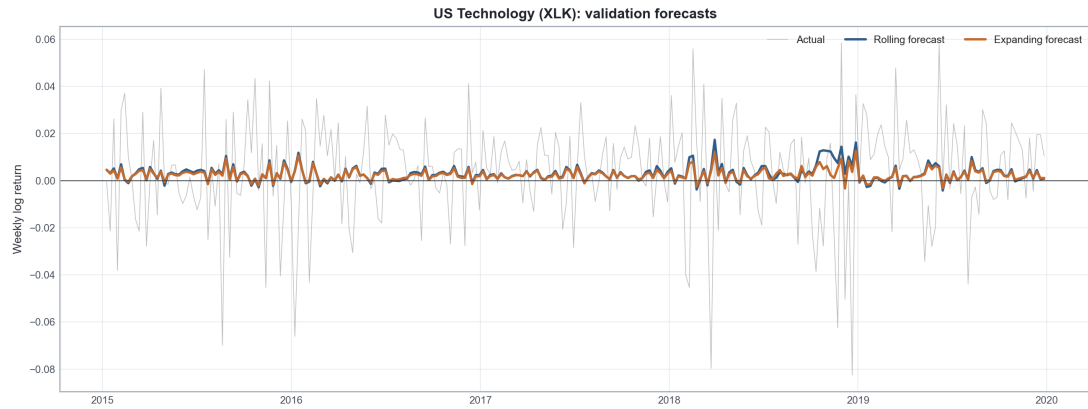
statistically significant; the only marked coefficient is the negative European rolling correlation, which is significant at the 10 percent level when the mean is tested across the eleven ETF-level correlations.

Table 5.1: Average forecast diagnostics in the validation period

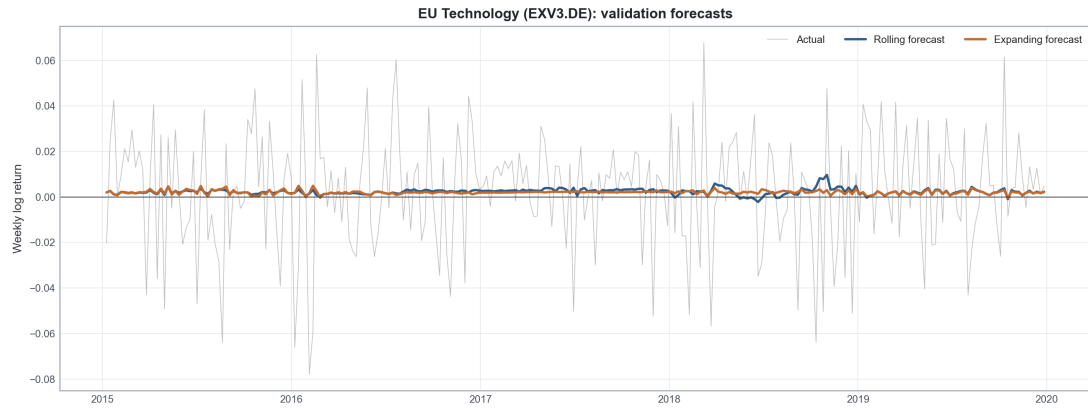
Market and window	Observations	RMSE	Forecast correlation	Hit ratio
US rolling	260	0.0214	-0.0197	0.5420
US expanding	260	0.0213	0.0201	0.5517
Europe rolling	260	0.0251	-0.0440*	0.5297
Europe expanding	260	0.0252	-0.0201	0.5336

Note: Forecast correlation is the mean of the eleven ETF-level Pearson correlations. The asterisk denotes significance of the mean correlation across ETFs at the 10 percent level; no coefficient is significant at the 5 percent level.

The hit ratios initially appear more favorable because all four values exceed 50 percent. This comparison is misleading, however, because positive returns represent 56.61 percent of the US observations and 53.64 percent of the European observations. A rule that always predicts a positive return would therefore attain a higher hit ratio than either econometric window in both regions. The forecasts themselves are strongly tilted toward positive signs and identify negative weeks poorly. The hit ratios should consequently be interpreted as evidence of the unconditional positive-return frequency rather than convincing directional predictability.



US technology ETF



European technology ETF

Figure 5.1: Rolling and expanding forecasts versus actual weekly log returns for the matched technology ETFs during validation.

The fitted forecasts in Figure 5.1 are small and close to the local mean, as expected for low-order econometric models applied to noisy weekly returns. This is true for both rolling and expanding estimation. The implication is that the forecast signal needs to be scaled carefully: alpha is important because it determines whether a statistically small forecast difference becomes an economically meaningful portfolio

tilt or remains too weak to affect allocation.

These conventional diagnostics are nevertheless not fully aligned with the allocation problem. RMSE evaluates the numerical accuracy of each ETF forecast through time, and the hit ratio evaluates whether its sign is correct. The 130/30 strategy instead uses the relative magnitude of volatility-adjusted forecasts to allocate the long and short books. A forecast can therefore be inaccurate as a point estimate while still being useful if ETFs receiving larger relative signals subsequently earn larger returns.

To test that channel without introducing another trading strategy, the volatility-adjusted signal is standardized across ETFs each week. Let

$$s_{i,t} = \frac{\hat{r}_{i,t}}{\hat{\sigma}_{i,t-1}}, \quad z_{i,t} = \frac{s_{i,t} - \bar{s}_t}{\text{sd}_i(s_{i,t})}. \quad (5.1)$$

The following cross-sectional regression is then estimated separately at every date:

$$r_{i,t} = a_t + b_t z_{i,t} + \varepsilon_{i,t}. \quad (5.2)$$

The intercept absorbs the common weekly return, while b_t measures whether an ETF with a signal one cross-sectional standard deviation higher realizes a higher return. Table 5.2 reports the time-series mean of the 260 weekly slopes, with Newey–West standard errors using four lags.

Table 5.2: Average weekly cross-sectional signal slopes

Market and window	Weeks	Mean slope (%)	HAC t -statistic	HAC p -value
US rolling	260	0.0490	1.491	0.136
US expanding	260	0.0519	1.501	0.133
Europe rolling	260	-0.0065	-0.170	0.865
Europe expanding	260	-0.0066	-0.170	0.865

The US slopes are positive and economically interpretable: a one-standard-deviation increase in the relative signal is associated with approximately 4.9–5.2 basis points of additional weekly return. The estimates are not statistically significant at the 10 percent level, however. The European slopes are slightly negative, close to zero, and clearly insignificant.

A complementary conditional-mean analysis provides a more transparent view of the same relation. At every date, ETFs are sorted into low, middle, and high terciles using $z_{i,t}$, after which the next realized return is averaged within each group. This procedure does not form a new portfolio and does not apply alpha, transaction costs, or 130/30 normalization. It simply asks whether realized returns increase monotonically with the signal used by the allocation rule. Table 5.3 reports the conditional means and tests the high-minus-low difference with Newey–West standard errors.

Table 5.3: Conditional weekly returns by volatility-adjusted signal tercile

Market and window	Low (%)	Middle (%)	High (%)	High minus Low (%)	HAC p -value
US rolling	0.1251	0.1369	0.2206	0.0955	0.1956
US expanding	0.1033	0.1674	0.2195	0.1161	0.1108
Europe rolling	0.1372	0.1118	0.1183	-0.0189	0.8373
Europe expanding	0.1155	0.1364	0.1215	0.0060	0.9523

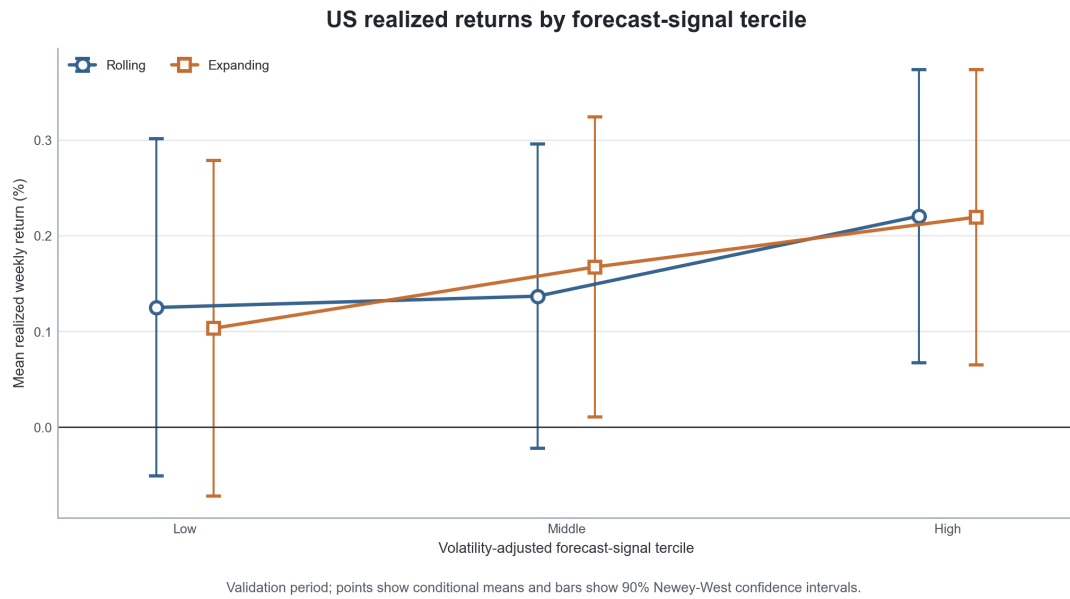


Figure 5.2: US realized returns conditional on the volatility-adjusted forecast-signal tercile. Error bars are 90 percent Newey–West confidence intervals.

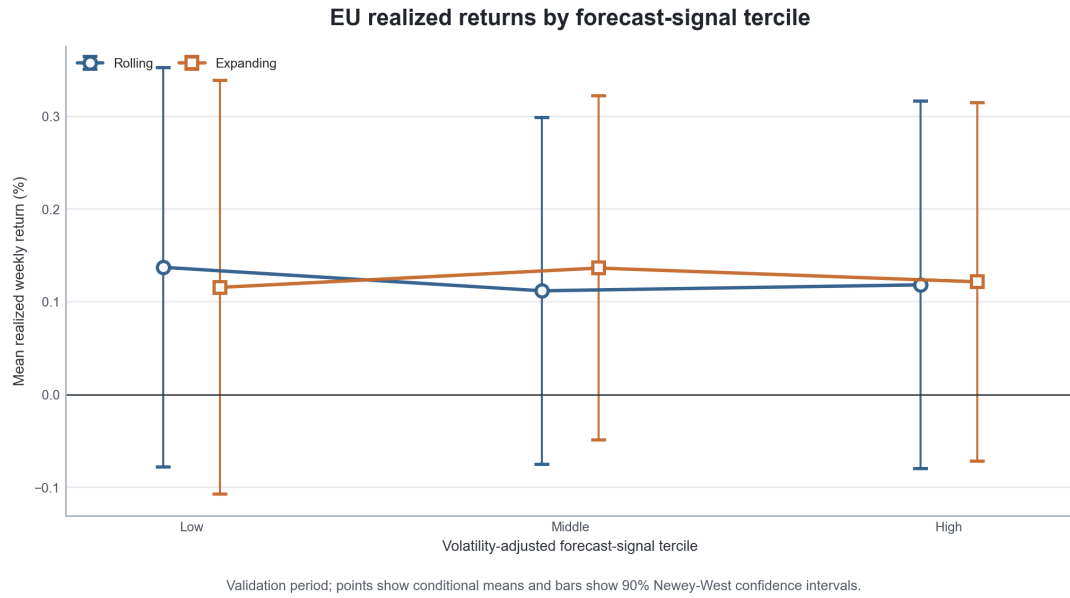


Figure 5.3: European realized returns conditional on the volatility-adjusted forecast-signal tercile. Error bars are 90 percent Newey–West confidence intervals.

Figure 5.2 shows a monotonic US pattern for both estimation windows. Mean realized returns increase from low to middle to high signals, and the high group earns approximately 0.22 percent per week under both specifications. The high-minus-low difference is 0.0955 percent for rolling estimation and 0.1161 percent for expanding estimation. Neither difference is significant at the 10 percent level; the expanding result, with $p = 0.1108$, narrowly misses that threshold. The wide confidence intervals therefore require a cautious interpretation, but the ordering is qualitatively consistent with the relative allocation mechanism.

Europe provides a clear contrast. Figure 5.3 contains no monotonic ordering under either window: the middle or low group can outperform the high

group, and both high-minus-low differences are economically small and statistically insignificant. The validation diagnostics therefore create a more favorable expectation for forecast-based portfolio construction in the US universe than in Europe. This is an expectation based on relative signal ordering rather than proof of portfolio performance; the independent test remains necessary to determine whether the pattern survives beyond validation.

5.2 Validation-period portfolio analysis

5.2.1 Portfolio weights and static alpha

Figure 5.4 illustrates how alpha changes the US allocation. A larger alpha makes the portfolio respond more strongly to the volatility-adjusted forecast signal. The effect is not limited to the speed at which existing positions are rebalanced: alpha can alter the subsequent trajectory of an ETF weight. This is visible for the Technology ETF after 2018. Under $\alpha = 0.2$, Technology remains on a comparatively persistent allocation path, whereas under $\alpha = 0.4$ its weight follows a markedly different adjustment. A more aggressive alpha therefore changes both the timing and the direction of sector exposures rather than simply accelerating an otherwise identical portfolio path.

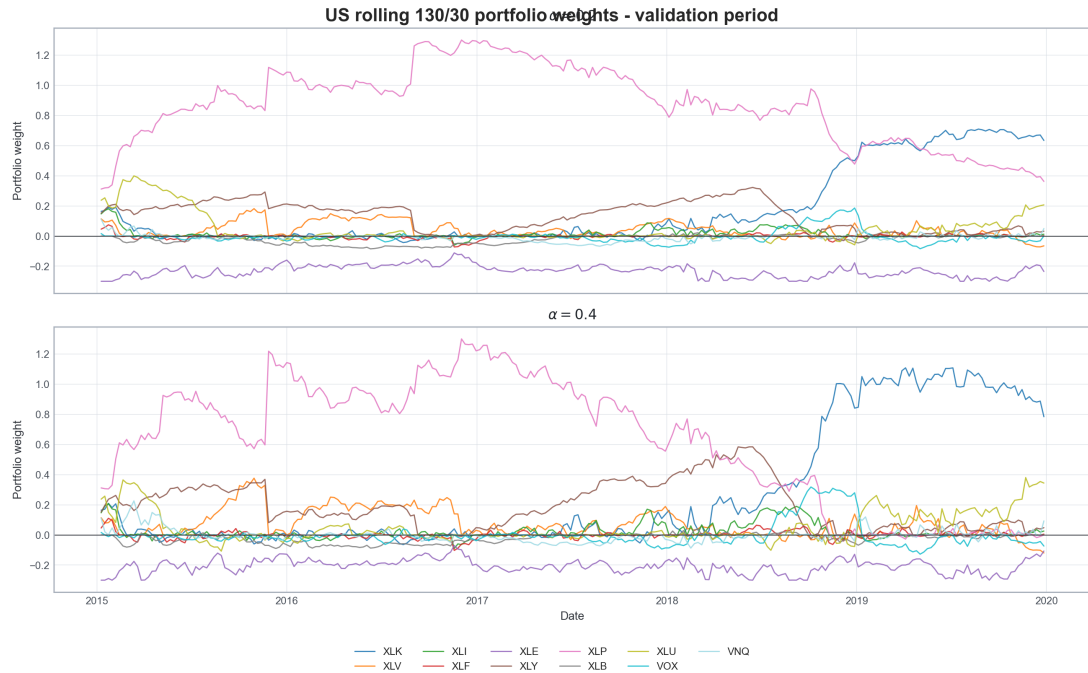


Figure 5.4: US rolling 130/30 portfolio weights for $\alpha = 0.2$ and $\alpha = 0.4$.

The complete static-alpha weight panels for both regions are reported in Appendix B. The comparison confirms that alpha is an aggressiveness parameter rather than a leverage parameter: all portfolios retain the same 130/30 exposure constraints, but higher alpha transmits forecast changes more rapidly into the long and short books.



Figure 5.5: Annualized return and Sharpe ratio across the static-alpha grid.

Figure 5.5 shows that the annualized-return and Sharpe-ratio curves are qualitatively very similar in both regions. Consequently, annualized volatility does not appear to change enough across alpha values to overturn the return ranking: most movements in Sharpe mirror movements in return. In the US, both measures display a clear maximum in the 0.2–0.3 region, followed by a sharp decline, a partial recovery at intermediate alpha values, and a renewed deterioration for large alpha. In Europe, a first favorable region appears around $\alpha = 0.05$, but it is dominated by the stronger Sharpe outcome at $\alpha = 1.00$; performance then declines as alpha increases further. The US Sharpe reaches its exact maximum at $\alpha = 0.30$, while the pre-specified parsimonious rule chooses the smallest alpha attaining at least 95 percent of that maximum. It therefore freezes $\alpha = 0.22$, whose Sharpe is 1.1328 compared with 1.1578 at $\alpha = 0.30$. In Europe, $\alpha = 1.00$ is both the exact Sharpe maximizer and the selected value.

Table 5.4: Static-alpha selection using validation data only

Market	Best alpha	Best Sharpe	Selected alpha	Selected Sharpe
US	0.30	1.1578	0.22	1.1328
Europe	1.00	0.6183	1.00	0.6183

5.2.2 Rolling versus expanding alpha rules

The internal selections remain walk-forward: the rolling rule uses the most recent 26 observations, while the expanding rule retains all eligible past observations. Figures 5.6 and 5.7 display the resulting alpha paths.

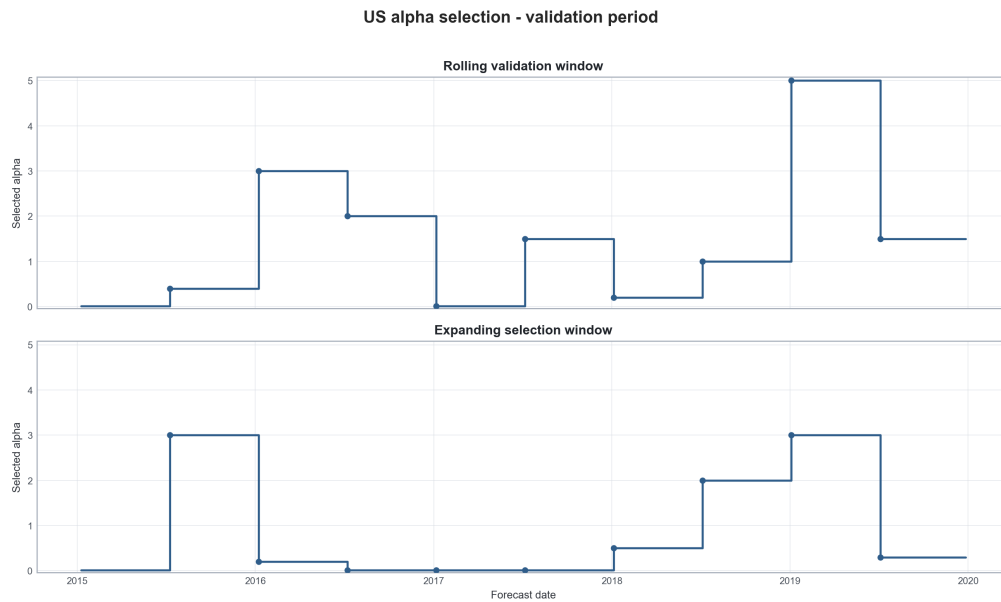


Figure 5.6: US rolling and expanding alpha-selection paths.

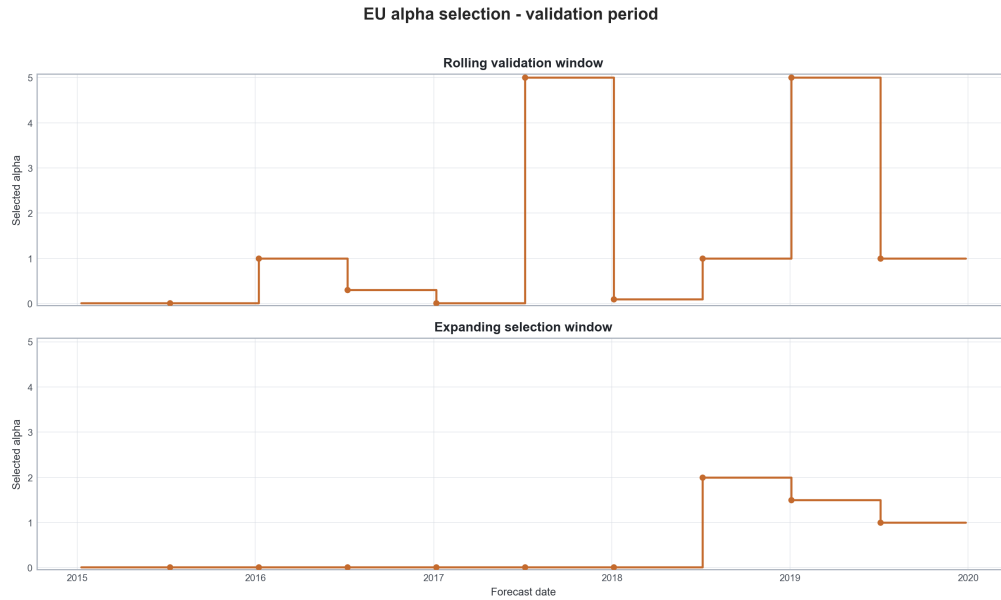


Figure 5.7: European rolling and expanding alpha-selection paths.

As expected, the expanding rule generally selects smaller alpha values and produces paths that are more stable over time. Because every new observation receives progressively less influence within a growing sample, isolated changes in recent performance rarely overturn the accumulated ranking. The rolling rule is more variable and occasionally moves to much larger alpha values because its fixed-length window gives recent regimes greater weight. Despite the greater stability of expanding selection, the results favor the adaptability of the rolling model in both regions. Its Sharpe is 0.8986 in the US and 0.6111 in Europe, compared with 0.7158 and 0.5531 for the corresponding expanding rules.

Table 5.5: Alpha-rule selection in the validation period

Market	Rule	Cum. return	Ann. return	Ann. vol.	Sharpe	Max drawdown
US	Rolling	0.7979	0.1245	0.1385	0.8986	-0.1775
US	Expanding	0.6213	0.1015	0.1418	0.7158	-0.1596
Europe	Rolling	0.5650	0.0937	0.1533	0.6111	-0.2215
Europe	Expanding	0.5044	0.0851	0.1539	0.5531	-0.2511

5.2.3 Mean-variance validation

The selected mean-variance window differs by region. Rolling mean-variance has the higher US Sharpe, 0.9514 versus 0.9354. In Europe, expanding mean-variance wins with Sharpe 0.5911 versus 0.4895 for rolling estimation. This regional difference would have been hidden by imposing one window choice on both markets.

Table 5.6: Mean-variance selection in the validation period

Market	Window	Cum. return	Ann. return	Ann. vol.	Sharpe	Max drawdown
US	Rolling	1.0034	0.1491	0.1567	0.9514	-0.1913
US	Expanding	0.8744	0.1339	0.1431	0.9354	-0.1595
Europe	Rolling	0.4994	0.0844	0.1724	0.4895	-0.2433
Europe	Expanding	0.5792	0.0957	0.1619	0.5911	-0.2325

5.2.4 Models frozen at the validation cutoff

Table 5.7 summarizes the only active specifications permitted to enter the final comparison. All choices are region-specific and are completed before any return from January 2020 onward is evaluated.

Table 5.7: Specifications frozen on December 31, 2019

Market	Static alpha	Alpha rule	MV 130/30
US	Rolling, $\alpha = 0.22$	Rolling	Rolling
Europe	Rolling, $\alpha = 1.00$	Rolling	Expanding

5.3 Transaction costs in validation

All active strategy returns subtract 5 basis points per unit of turnover. Figure 5.8 shows that implementation cost increases with alpha because more aggressive signals produce larger reallocations.

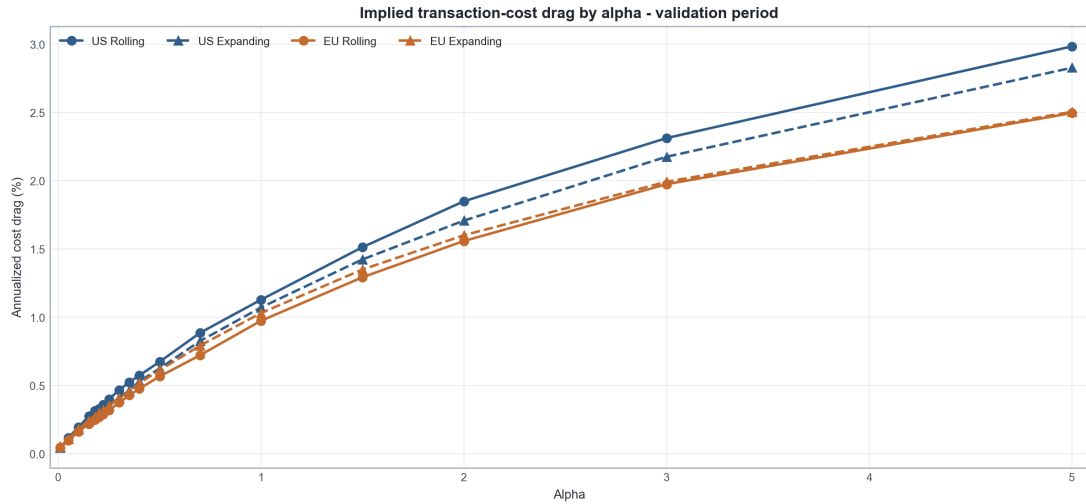


Figure 5.8: Implied annualized transaction-cost drag across the static-alpha grid.

The curves are monotonically increasing but distinctly concave. Cost drag rises steeply for small and intermediate alpha values because an initial increase in signal aggressiveness moves many portfolio weights away from their previous allocations. The slope then decreases: once the 130/30 normalization has pushed assets firmly toward the long or short side, an additional increase in alpha changes the normalized allocation by progressively less. This creates a saturation effect at large alpha values. The US rolling curve is generally the highest, consistent with larger and more frequent reallocations, while the European rolling and expanding curves become increasingly similar toward the upper end of the grid. The interior US selection is therefore consistent with a balance between exploiting the forecast signal and avoiding the flatter, high-cost region in which additional aggressiveness produces limited allocation benefit.

5.4 True out-of-sample strategy comparison

The final test contains 333 weekly observations from January 3, 2020 through May 15, 2026. Only the frozen specifications in Table 5.7 are retained. The benchmark and scarecrow require no test-period optimization. Portfolio weights are not reset at the cutoff, but every performance statistic below uses only returns strictly after December 31, 2019.

5.4.1 United States

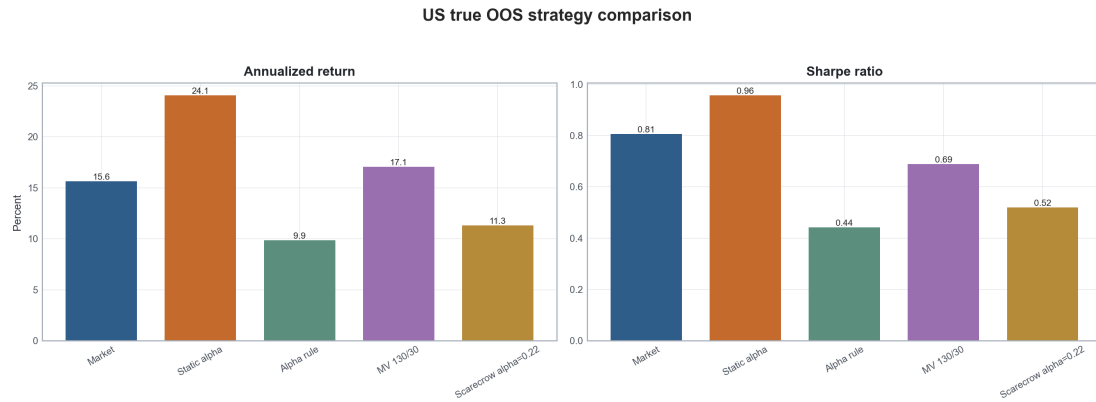


Figure 5.9: US true OOS annualized return and Sharpe-ratio comparison.

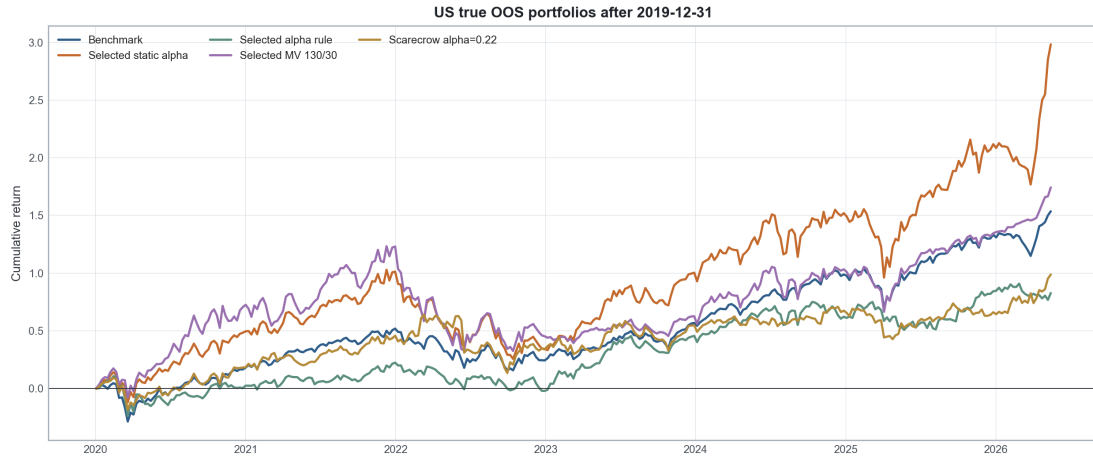


Figure 5.10: US true OOS cumulative portfolio paths.

Table 5.8: Final US true out-of-sample results

Portfolio	Cum. return	Ann. return	Ann. vol.	Sharpe	Max drawdown	Turnover
Market benchmark	1.5357	0.1564	0.1938	0.8071	-0.3183	—
Static $\alpha = 0.22$	2.9832	0.2409	0.2515	0.9578	-0.3809	0.1741
Rolling alpha rule	0.8267	0.0987	0.2232	0.4420	-0.3135	0.4369
Rolling MV 130/30	1.7427	0.1706	0.2476	0.6892	-0.4067	0.0285
Scarecrow ($\alpha = 0.22$)	0.9870	0.1132	0.2179	0.5194	-0.3079	0.8867

The selected static-alpha strategy is the strongest US specification. It earns 24.09 percent annualized with Sharpe 0.9578, compared with 15.64 percent and 0.8071 for the market. The selected rolling mean-variance portfolio improves annualized return relative to the market but has lower risk-adjusted performance and a deeper drawdown. The rolling alpha rule does not preserve its validation advantage and is materially below the market. The scarecrow, evaluated at the same $\alpha = 0.22$ as

the selected econometric strategy, improves on the rolling alpha rule but remains far below the static econometric portfolio. Holding alpha fixed therefore makes the incremental contribution of the econometric forecast signal particularly clear.

5.4.2 Europe

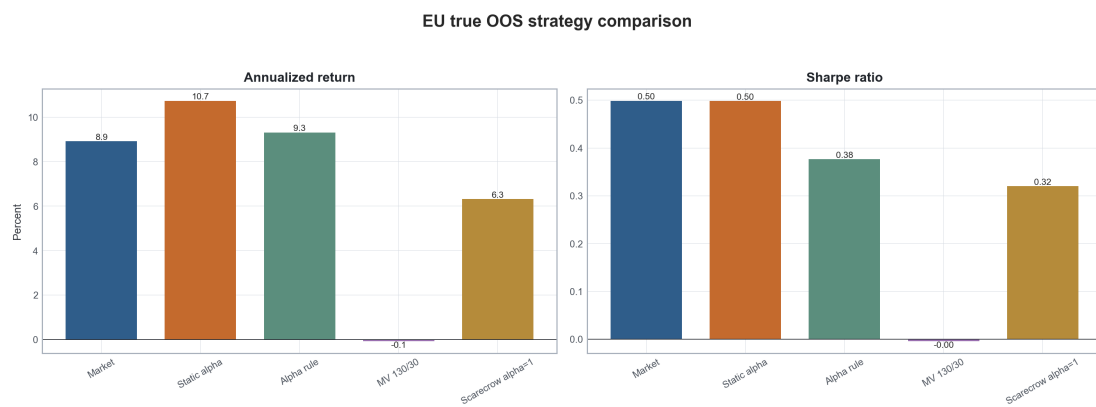


Figure 5.11: European true OOS annualized return and Sharpe-ratio comparison.

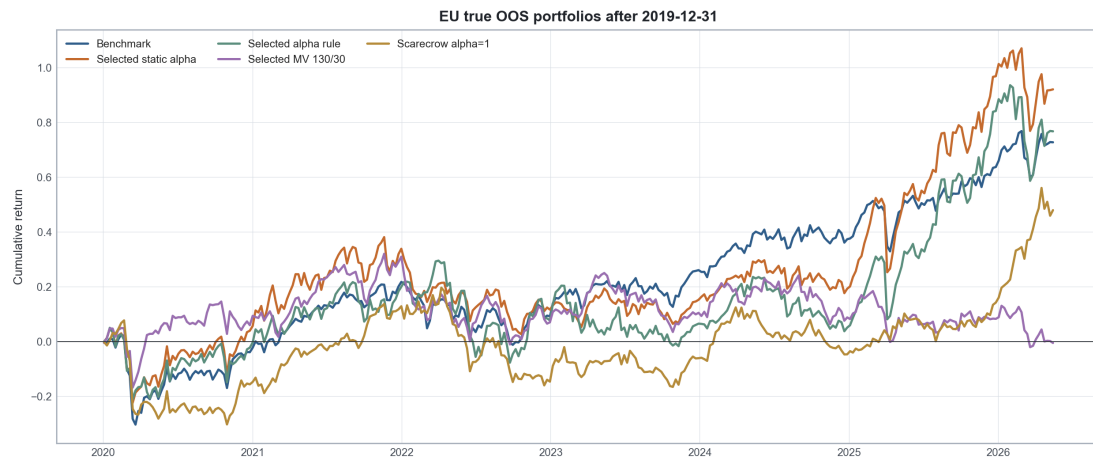


Figure 5.12: European true OOS cumulative portfolio paths.

Table 5.9: Final European true out-of-sample results

Portfolio	Cum. return	Ann. return	Ann. vol.	Sharpe	Max drawdown	Turnover
Market benchmark	0.7279	0.0892	0.1788	0.4986	-0.3205	–
Static $\alpha = 1.00$	0.9212	0.1073	0.2155	0.4981	-0.2584	0.3986
Rolling alpha rule	0.7680	0.0931	0.2472	0.3765	-0.2876	0.4181
Expanding MV 130/30	-0.0049	-0.0008	0.1829	-0.0042	-0.2581	0.0158
Scarecrow ($\alpha = 1.00$)	0.4801	0.0631	0.1973	0.3200	-0.3525	1.8941

The selected European static-alpha portfolio has the highest cumulative and annualized return and a smaller maximum drawdown than the market. Its higher volatility leaves its Sharpe ratio, 0.4981, almost identical to the market's 0.4986. The rolling alpha rule slightly exceeds the market in cumulative and annualized return but has a lower Sharpe ratio. The expanding mean-variance specification, despite winning in validation, is approximately flat in the test period. At the same $\alpha = 1.00$ used by the selected econometric strategy, the scarecrow remains below both the static portfolio and the rolling alpha rule on return and Sharpe ratio, while generating substantially higher turnover.

Chapter 6

Discussion

The selected static-alpha portfolio remains robust in the US test and generates a substantial improvement over the market. The European static strategy improves return and drawdown but not risk-adjusted performance. Dynamic alpha selection and mean-variance allocation are much less stable across the cutoff.

The static-alpha results are consistent with the idea that weak forecasts can be economically useful after cross-sectional scaling and constrained portfolio mapping. The US result is particularly strong: the validation rule deliberately chooses $\alpha = 0.22$ rather than the exact in-sample maximum at $\alpha = 0.30$, yet the frozen value performs well in the independent test. This reduces concern that the result depends on selecting a single noisy grid maximum. Europe provides a more cautious result because the selected strategy's return advantage is offset by higher volatility.

The time-varying alpha rules illustrate the difference between adaptive design and realized robustness. Rolling selection wins validation in both regions, but its test-period Sharpe falls to 0.4420 in the US and 0.3765 in Europe. Continual reselection from recent observations can respond to regime changes, but it can also amplify transient performance rankings. Flexibility alone is therefore not evidence of superior OOS performance.

The mean-variance results reinforce this point. Validation selects different estimation windows across regions, demonstrating the value of region-specific model choice. Nevertheless, the US rolling mean-variance portfolio does not beat the market on Sharpe in the test, and the European expanding portfolio is approximately flat. Classical optimization remains sensitive to expected-return and covariance estimation error even when exposure constraints and a covariance ridge are imposed.

Matching the scarecrow alpha to the selected econometric alpha sharpens the interpretation of the independent test. In the US, the econometric and scarecrow portfolios use the same $\alpha = 0.22$, yet the selected econometric strategy delivers materially higher return and Sharpe ratio. The same ordering appears in Europe at $\alpha = 1.00$: the econometric static strategy outperforms the scarecrow on return, Sharpe ratio, drawdown, and turnover. Since volatility scaling, exposure constraints, transaction costs, and signal aggressiveness are held constant within each region, these differences provide more direct evidence that the econometric model contributes useful information beyond the naive lagged-return signal.

Transaction costs and downside risk remain important qualifications. All active returns are net of the 5 basis point turnover charge, and the final tables report turnover observed in the test period. The static US strategy's higher return is accompanied by higher volatility and a larger drawdown than the market. The results should therefore be read as a risk-return comparison rather than as unconditional dominance.

The test includes the COVID shock and the subsequent inflation and monetary-policy regime. These observations were never used to select the frozen specifications, making them a demanding evaluation of robustness. The sharp drawdowns and reversals in Figures 5.10 and 5.12 show that low-order linear forecasts remain vulnerable during abrupt repricing even when their longer-run portfolio performance is useful. At the same time, placing the COVID episode entirely inside the test strengthens the external validity of the evidence: the selected portfolios are evaluated not only in ordinary conditions but also through an unforeseen period of extreme volatility and market disruption. In particular, the persistence of the static US strategy's advantage after this shock supports its robustness beyond relatively stable regimes.

Chapter 7

Conclusion and Future Research

This thesis evaluates BIC-selected AR, MA, and ARMA forecasts for 130/30 allocation across matched US and European sector ETF universes. Its central methodological feature is a strict separation between validation and final testing. Static alpha, rolling versus expanding alpha selection, and rolling versus expanding mean-variance allocation are chosen using observations no later than December 31, 2019. Only those frozen choices enter the 333-week true OOS period beginning January 3, 2020.

The strongest evidence is obtained for the US static-alpha portfolio. The validation procedure selects $\alpha = 0.22$, and the strategy subsequently achieves annualized return of 24.09 percent and Sharpe ratio 0.9578, compared with 15.64 percent and 0.8071 for the S&P 500 benchmark. In Europe, the selected $\alpha = 1.00$ portfolio raises annualized return from 8.92 to 10.73 percent and reduces maximum drawdown, but its Sharpe ratio is essentially unchanged relative to the market.

The remaining families are less robust. The rolling alpha rule wins validation in both regions but underperforms the market on test-period Sharpe. Rolling mean-variance is selected for the US and expanding mean-variance for Europe; neither beats its market benchmark on test-period Sharpe, and the European

portfolio is approximately flat. These reversals are not incidental: they demonstrate the importance of reporting validation selection separately from genuine OOS performance.

The findings support a qualified conclusion. Econometric forecasts can contain economically useful cross-sectional information, especially when combined with volatility scaling, transaction costs, and disciplined exposure constraints. That value is strategy- and region-dependent, however, and is most convincing for the static US specification. Adaptive rules should not be assumed to generalize simply because they perform well during validation.

Future research should repeat the same frozen-selection design across alternative cutoff dates and multiple non-overlapping test periods. Other extensions include different validation lengths, explicit turnover penalties, covariance shrinkage, position caps, time-varying trading costs, GARCH-type volatility forecasts, nonlinear or regime-switching return models, and broader asset universes. Repeated true OOS evaluation would help determine whether the selected alpha regions reflect persistent economic structure or one historical regime.

Appendix A: ACF Diagnostics

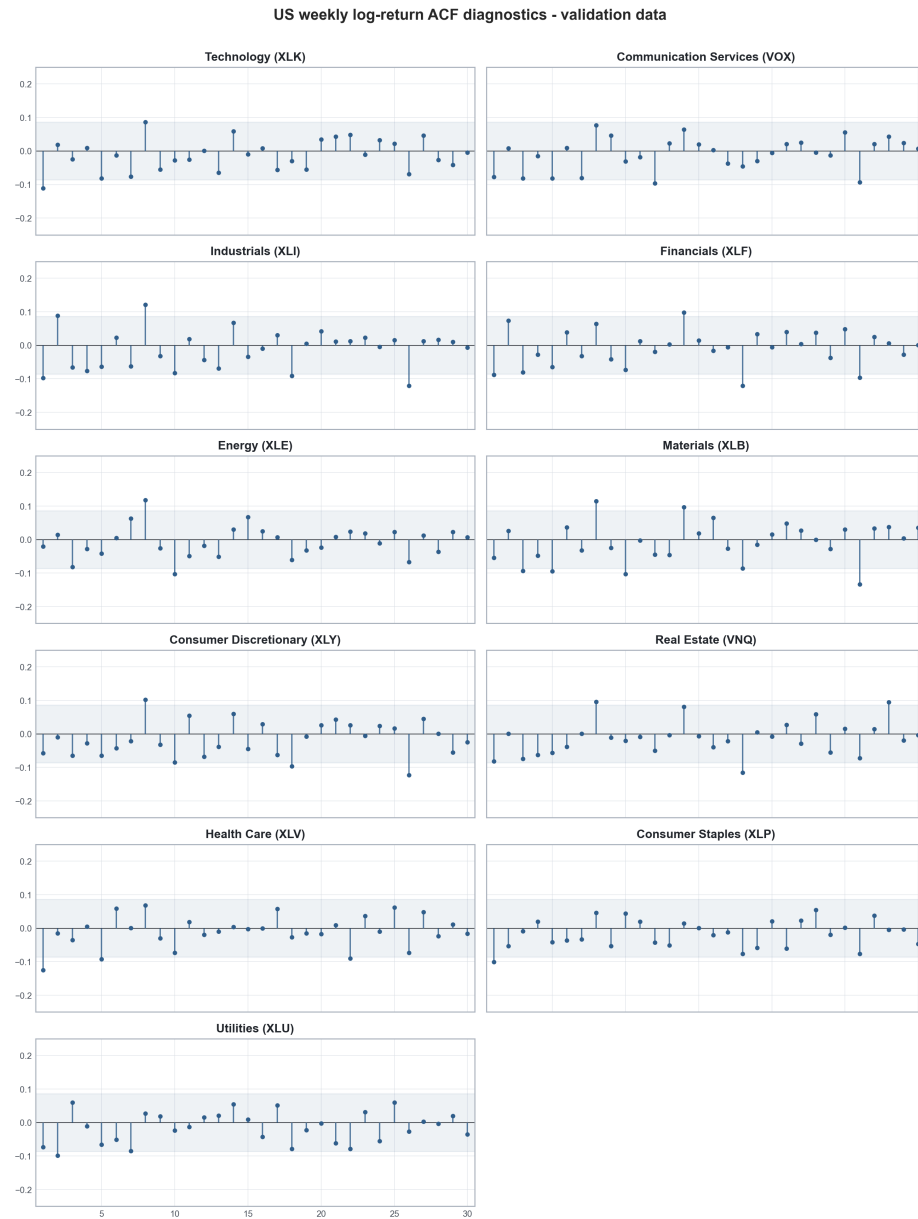


Figure A.1: Complete US weekly log-return ACF diagnostics.

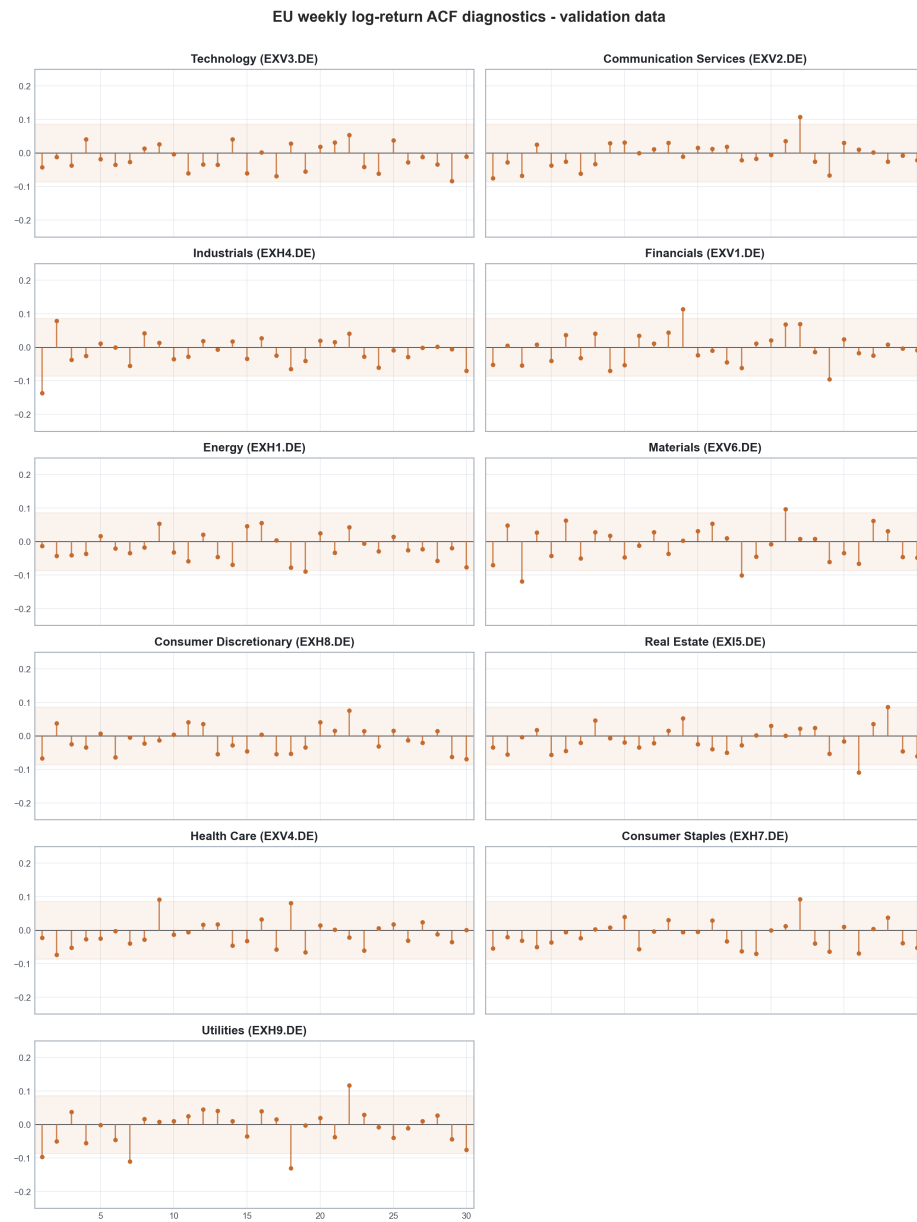


Figure A.2: Complete European weekly log-return ACF diagnostics.

Appendix B: Portfolio Weights

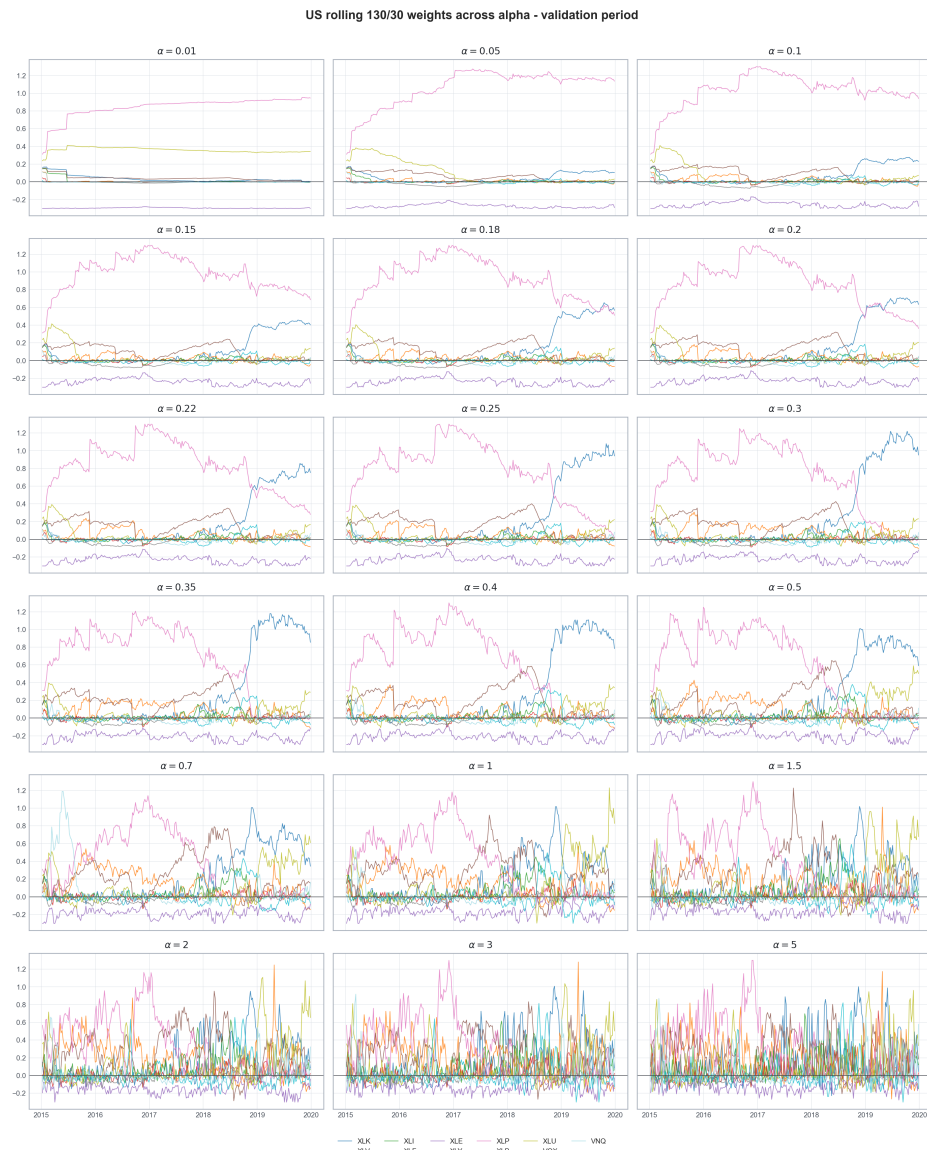


Figure B.1: US rolling 130/30 portfolio weights across the static alpha grid.

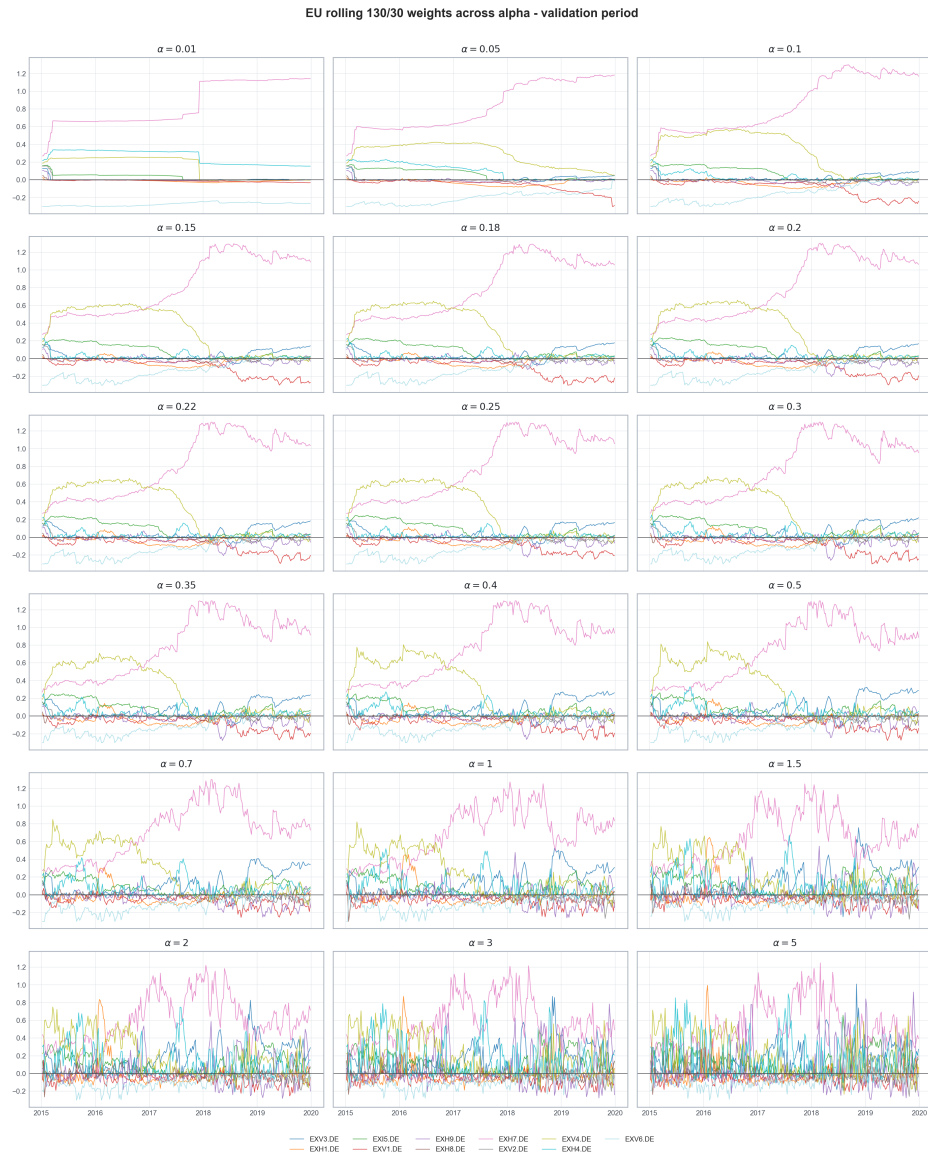


Figure B.2: European rolling 130/30 portfolio weights across the static alpha grid.

Bibliography

- [1] Box, G. E. P., and Jenkins, G. M. (1970). *Time Series Analysis: Forecasting and Control*. Holden-Day.
- [2] Tsay, R. S. (2010). *Analysis of Financial Time Series*, 3rd ed. Wiley.
- [3] Campbell, J. Y., Lo, A. W., and MacKinlay, A. C. (1997). *The Econometrics of Financial Markets*. Princeton University Press.
- [4] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464.
- [5] Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, 7(1), 77–91.
- [6] Markowitz, H. M. (1959). *Portfolio Selection: Efficient Diversification of Investments*. John Wiley & Sons.
- [7] Merton, R. C. (1972). An analytic derivation of the efficient portfolio frontier. *Journal of Financial and Quantitative Analysis*, 7(4), 1851–1872.
- [8] Michaud, R. O. (1989). The Markowitz optimization enigma: Is “optimized” optimal? *Financial Analysts Journal*, 45(1), 31–42.
- [9] Jagannathan, R., and Ma, T. (2003). Risk reduction in large portfolios: Why imposing the wrong constraints helps. *The Journal of Finance*, 58(4), 1651–1683.

- [10] Welch, I., and Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, 21(4), 1455–1508.
- [11] Goyal, A., Welch, I., and Zafirov, A. (2024). A comprehensive 2022 look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, 37(11), 3490–3557.
- [12] Campbell, J. Y., and Thompson, S. B. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *The Review of Financial Studies*, 21(4), 1509–1531.
- [13] Rapach, D. E., Strauss, J. K., and Zhou, G. (2010). Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. *The Review of Financial Studies*, 23(2), 821–862.
- [14] Granger, C. W. J., and Pesaran, M. H. (2000). Economic and statistical measures of forecast accuracy. *Journal of Forecasting*, 19(7), 537–560.
- [15] Cenesizoglu, T., and Timmermann, A. (2012). Do return prediction models add economic value? *Journal of Banking & Finance*, 36(11), 2974–2987.
- [16] Moskowitz, T. J., and Grinblatt, M. (1999). Do industries explain momentum? *The Journal of Finance*, 54(4), 1249–1290.

- [17] Beber, A., Brandt, M. W., and Kavajecz, K. A. (2011). What does equity sector orderflow tell us about the economy? *The Review of Financial Studies*, 24(11), 3688–3730.
- [18] Rapach, D. E., Strauss, J. K., Tu, J., and Zhou, G. (2019). Industry return predictability: A machine learning approach. *The Journal of Financial Data Science*, 1(3), 9–28.
- [19] Gu, S., Kelly, B., and Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5), 2223–2273.
- [20] Bessler, W., and Wolff, D. (2024). Portfolio optimization with sector return prediction models. *Journal of Risk and Financial Management*, 17(6), 254.
- [21] DeMiguel, V., Garlappi, L., and Uppal, R. (2009). Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy? *The Review of Financial Studies*, 22(5), 1915–1953.
- [22] Lo, A. W., and Patel, P. N. (2008). 130/30: The new long-only. *The Journal of Portfolio Management*, 34(2), 12–38.
- [23] White, H. (2000). A reality check for data snooping. *Econometrica*, 68(5), 1097–1126.
- [24] Frazzini, A., Israel, R., and Moskowitz, T. J. (2018). Trading costs. *Available at SSRN*.

- [25] S&P Dow Jones Indices. S&P U.S. Indices Methodology and S&P 500 index overview. <https://www.spglobal.com/spdji/en/indices/equity/sp-500/>.
- [26] STOXX Ltd. (2024). STOXX Europe 600 and Derived Indices: Index Construction. https://stoxx.com/wp-content/uploads/2024/06/STOXX_Europe_600_202406.pdf.
- [27] State Street Global Advisors. State Street Communication Services Select Sector SPDR ETF (XLC), fund page. <https://www.ssga.com/us/en/individual/etfs/funds/the-communication-services-select-sector-spdr-fund-xlc>.
- [28] State Street Global Advisors. State Street Real Estate Select Sector SPDR ETF (XLRE), fund page. <https://www.ssga.com/us/en/intermediary/etfs/the-real-estate-select-sector-spdr-fund-xlre>.
- [29] Vanguard. Vanguard Communication Services ETF (VOX), fund page. <https://advisors.vanguard.com/investments/products/vox/vanguard-communication-services-etf>.
- [30] Vanguard. Vanguard Real Estate ETF (VNQ), fund page. <https://investor.vanguard.com/investment-products/etfs/profile/vnq>.
- [31] BlackRock. iShares STOXX Europe 600 UCITS ETF (DE) (EXSA), fund page. <https://www.blackrock.com/ch/individual/en/products/251931/ishares-stoxx-europe-600-ucits-etf-de>.

- [32] BlackRock. iShares STOXX Europe 600 Technology UCITS ETF (DE) (EXV3), fund page. <https://www.blackrock.com/uk/individual/products/251961/ishares-stoxx-europe-600-technology-ucits-etf-de-fund>.