UNIVERSITAT D'ALACANT
UNIVERSIDAD DE ALICANTE
Facultat de Ciències Econòmiques i Empresarials
Facultad de Ciencias Económicas y Empresariales

MÁSTER EN ECONOMICS
WITH DATA SCIENCE

CURSO ACADÉMICO 2025 - 2026

# DO SMART BETA ETFs EXHIBIT MARKET TIMING?

## STATISTICAL EVIDENCE AND ECONOMIC VALUE

PEDRO MIGUEL SAPENA LÓPEZ

ÁNGEL MANUEL LEÓN VALLE

DEPARTAMENTO DE FUNDAMENTOS
DEL ANÁLISIS ECONÓMICO

San Vicente del Raspeig, septiembre de 2026

# Do Smart Beta ETFs Exhibit Market Timing?
# Statistical Evidence and Economic Value

Master's Thesis

**Abstract**

We test for market and volatility timing in a sample of eighty US smart beta ETFs over the 2013–2026 period and assess whether any such evidence can be economically exploited. We estimate twenty-three specifications combining factor models with the Treynor-Mazuy and Henriksson-Merton tests and a volatility timing term, at daily and monthly frequency, complemented by Markov regime-switching models. We find robust negative abnormal returns of between 2.8% and 5.2% annually, most pronounced among value and dividend strategies. Timing tests yield systematically negative coefficients, indicating unfavourable exposure. This apparent evidence proves unstable across specifications and subperiods, cannot be cleanly separated from volatility exposure, and does not survive multiple testing correction through a block bootstrap. Consistently, no strategy built on these signals outperforms a passive position out of sample.

# Contents

# 1 Introduction

The exchange-traded fund industry has changed substantially over the past two decades, and not only in size. Alongside conventional index products that track capitalization-weighted benchmarks, a middle category has emerged, commonly labelled *smart beta*: funds that track indices built on selection and weighting rules which deliberately depart from market capitalization in order to capture specific risk premia or accounting characteristics.

Before stating the question that motivates this study, we define the concept at its core. Market timing refers to the ability to adjust a portfolio's market exposure according to expected market movements, raising it ahead of advances and lowering it ahead of declines. A manager with such ability would enjoy a considerable advantage, participating in upswings while avoiding at least part of the drawdowns.

What makes this ability empirically tractable is that it leaves an observable trace. If market exposure varies systematically, portfolio sensitivity is no longer constant and the relationship between fund and market returns becomes non-linear. Timing can therefore be tested using return series alone, without observing the manager's actual decisions, which are rarely available. Alongside directional timing, the literature has examined a second variety, volatility timing, which concerns not whether the market rises or falls but how violently it moves. We test for both.

One terminological clarification is essential before proceeding. Throughout this thesis, timing refers to return patterns identified by standard market-timing tests, not to discretionary forecasting ability on the part of ETF managers. What we test is whether smart beta ETFs exhibit return non-linearities that conventional timing tests would classify as timing, a question that is well posed precisely because these products involve no discretionary decisions at all.

Smart beta products occupy an unusual position in the active-passive debate. No manager makes discretionary decisions about portfolio composition, which brings them close to passive management; yet they do not simply reproduce the market either, since their construction rules embed an implicit investment strategy. This hybrid nature raises a question the literature has not addressed systematically: do these rules generate, unintentionally, the kind of time-varying exposure that fund evaluation studies identify as market timing?

Our objective is twofold. We first test whether market and volatility timing are present in a sample of US smart beta ETFs. Conditional on that evidence, we then assess whether it can be exploited through an investment strategy implementable in real time. Four specific questions follow. Do these products deliver abnormal returns once we control for their exposure to standard risk factors? Do their returns display the non-linearity that timing tests are designed to detect? Does this behaviour differ across investment styles and across market states? And does whatever evidence we find survive rigorous statistical testing and an out-of-sample economic evaluation?

We analyse eighty US equity ETFs over the 2013–2026 period, at daily and monthly frequency. Our empirical strategy proceeds through increasing levels of stringency. We begin with conventional factor models and the classic timing tests of Treynor and Mazuy (1966) and Henriksson and Merton (1981), estimated over four alternative factor sets and augmented with a volatility timing term. We then disaggregate by smart beta family and estimate a Markov regime-switching model that allows coefficients to differ between bull and bear market states.

The resulting evidence is then subjected to three filters. The first concerns robustness: we test alternative specifications of market volatility and examine whether results hold across subperiods. The second addresses the multiple testing problem through a block bootstrap procedure, following Kosowski *et al.* (2006) and Fama and French (2010). The third is an out-of-sample evaluation that strictly respects the information available at each point in time and accounts for transaction costs.

Four findings emerge. First, the ETFs in our sample deliver negative and statistically significant abnormal returns, between 2.8% and 5.2% annually, and most pronounced among value and dividend strategies. This result holds across specifications, across subperiods, and under multiple testing correction. Second, conventional inference points to seemingly robust evidence of unfavourable market and volatility timing, affecting between fifteen and twenty-seven products depending on the specification. Third, this evidence does not survive multiple testing correction: the bootstrap reduces the number of significant coefficients to zero or one, which means the apparent timing is attributable to chance arising from testing eighty hypotheses at once. Fourth, and consistently, no timing strategy built on these signals outperforms a passive position out of sample, even before transaction costs.

We make three contributions. The first is substantive. Existing work has examined the factor exposures of smart beta products, the time variation of betas in managed portfolios, and market timing in actively managed funds, but these strands have remained largely separate. What has not been examined systematically is whether rules-based products, in which no discretionary timing decision is taken, nonetheless generate the kind of time-varying exposure that timing tests detect. That is the specific gap we address. The second is methodological: we document how severely conventional inference misleads in a multiple testing setting, showing that the common practice of counting how many assets display significant coefficients can produce entirely wrong conclusions.

The third concerns research design: we illustrate why statistical evidence must be subjected to an economic evaluation that respects real-time information, avoiding the look-ahead bias introduced by smoothed probabilities or contemporaneous regressors.

The paper proceeds as follows. Section 2 reviews the relevant literature. Section 3 describes the sample, the classification of products and the series we construct. Section 4 sets out the methodology. Section 5 presents the empirical results. Section 6 discusses their economic interpretation, notes the limitations of the analysis and concludes.

# 2 Literature review

## 2.1 Portfolio performance evaluation

Risk-adjusted performance measurement originates with Sharpe (1966) and, above all, Jensen (1968), who proposed measuring a portfolio's abnormal return as the intercept of a regression of its excess return on the market excess return. Jensen's alpha remains the benchmark measure, although the single-factor model soon proved inadequate.

Fama and French (1993) showed that size and book-to-market capture systematic variation in returns that the market model leaves unexplained, and proposed their three-factor model. Carhart (1997) added a momentum factor, motivated by the persistence documented in Jegadeesh and Titman (1993). More recently, Fama and French (2015) extended the specification with profitability and investment factors.

This progression matters directly for our purposes: estimated alphas depend critically on the factor set employed, which is why we run our analysis across several alternative specifications.

## 2.2 Market timing

The timing literature starts from the premise that a manager able to forecast market movements will adjust systematic risk exposure accordingly, generating a non-linear relationship between fund and market returns.

Treynor and Mazuy (1966) formalised this intuition by adding a quadratic term to the market regression; applying it to fifty-seven funds, they found timing ability in one. Henriksson and Merton (1981) developed an alternative in which exposure takes two values depending on the sign of the market return, and Henriksson (1984) applied the test to one hundred and sixteen funds, finding favourable evidence in three.

Subsequent evidence has been largely negative. Ferson and Schadt (1996) argued that unconditional tests may confound exposure shifts driven by public information with genuine ability, and proposed conditional versions. Becker *et al.* (1999), Goetzmann *et al.* (2000) and Jiang (2003) report similarly sparse findings.

One strand of work highlights the limitations of return-based tests. Jiang *et al.* (2007) argue that such tests suffer from an artificial timing bias driven by portfolio dynamics themselves, and develop a holdings-based measure instead. Bollen and Busse (2001) show that daily data materially change the conclusions relative to monthly data, since exposure decisions are taken at intervals shorter than a month. This consideration motivates our use of both frequencies.

## 2.3 Volatility timing

Beyond directional timing, the literature has explored whether managers adjust exposure in response to expected market volatility. Busse (1999) documents this using daily data, and later work has added a volatility term to the classic Treynor-Mazuy and Henriksson-Merton specifications.

Measuring market volatility admits two main approaches. The first models it through GARCH-type specifications (Bollerslev, 1986), which estimate conditional variance from the persistence of the series. The second uses realised volatility, the sum of squared higher-frequency returns within each period. Andersen and Bollerslev (1998) show the latter to be a considerably more precise estimator of actual volatility, since it exploits intra-period information. We compare both and document how sensitive our results are to this choice.

## 2.4 Luck, skill and multiple testing

A central concern in the recent literature is distinguishing genuine skill from chance. Lo and MacKinlay (1990) warn about the biases that arise from repeatedly mining the same data, and Sullivan *et al.* (1999) develop procedures to correct them.

In the fund evaluation context specifically, Kosowski *et al.* (2006) propose a bootstrap procedure that constructs the distribution of test statistics under the null of no skill, recognising that when a large number of funds is examined some will display extreme results by chance alone. Fama and French (2010) apply a comparable methodology and conclude that aggregate evidence of skill in the US fund industry is scarce once luck is accounted for.

This strand is directly relevant here, since we test eighty assets simultaneously across multiple specifications. Applying a block bootstrap, which additionally preserves the temporal dependence of the series, is a central element of our empirical strategy.

### 2.5 Regime-switching models

Markov regime-switching models, introduced into economic analysis by Hamilton (1989), allow the moments of a series to vary across unobservable latent states governed by a Markov chain. Financial applications are extensive: Ang and Bekaert (2002) use them for asset allocation, Perez-Quiros and Timmermann (2000) to study cyclical variation in returns, and Guidolin and Timmermann (2008) to analyse size and value anomalies.

Within fund evaluation, Kosowski (2006) documents that risk-adjusted alphas differ between expansions and recessions. Turtle and Zhang (2012) develop a multivariate regime-switching model for international funds and find that emerging market fund alphas are significantly positive in global bull regimes, and that betas are higher in bear regimes, an asymmetry they attribute to rising correlations during periods of stress.

We adopt their identification logic, estimating the regime on the market and applying the resulting state calendar to all assets, though we implement it in a univariate setting given the large number of assets involved.

### 2.6 Smart beta and factor investing

The growth of smart beta products responds to the empirical finding that certain security characteristics are persistently associated with return differentials. These products aim to deliver systematic exposure to such premia through transparent rules and at fees below those of traditional active management.

The literature has largely examined whether these products effectively capture the premia they target, with mixed results. Whether their construction rules generate time-varying exposure comparable to market timing has received little attention, and it is precisely this gap that our work addresses.

## 3 Data

### 3.1 ETF sample

Our sample consists of eighty US equity ETFs covering the period from 3 January 2013 to 2 March 2026. At daily frequency this yields 3,309 observations per asset; aggregated to monthly frequency, the sample spans 158 months (January 2013 to February 2026).

Two criteria guided this choice. We restrict the analysis to the US market to avoid the heterogeneity that arises from pooling markets with different regulatory regimes, trading hours and currencies, and because the benchmark risk factors are constructed on that market. We also require a complete return history over the whole period. This yields a balanced panel and avoids changes in cross-sectional composition over time, but it does not eliminate survivorship bias in a broad sense: it restricts the sample to long-lived products, excluding both later entrants and funds that ceased to exist, and may therefore introduce survivor-related selection.

## 3.2 What kind of products these are

It is worth being precise about the nature of the vehicles we study, since their interpretation hinges on it. The label "passive management" is often applied as if it described a homogeneous category, when in fact it spans products with very different degrees of intervention in portfolio construction.

A pure index fund tracks a capitalization-weighted benchmark and simply reproduces the market. Its systematic risk exposure is constant by construction, and nothing resembling market timing should be expected from it.

A smart beta or rules-based ETF also tracks an index, but one built on selection and weighting rules that deliberately depart from market capitalization. These products screen securities on valuation, dividend yield, accounting quality, historical volatility or momentum, and rebalance at predetermined intervals. No manager makes discretionary decisions, yet an investment strategy is embedded in the index rules.

Finally, some products approach what might be called systematic active management: selection depends on proprietary assessments by the index provider, or portfolio turnover is substantially higher.

Our sample consists predominantly of the second category, and this matters for the research question. Market timing in the traditional sense, understood as a manager's ability to anticipate market movements and adjust exposure accordingly, is not something we should expect to find. What is plausible, however, is that index construction rules generate *mechanically* a market exposure that varies over time and that timing tests will detect as such. A minimum volatility ETF, for instance, tilts away from more volatile securities as market volatility rises, not because anyone decides so, but because the index methodology dictates it. Likewise, a momentum index rotates towards recent winners, generating procyclical exposure.

Any timing evidence we obtain must therefore be read as a by-product of index construction rules, not as managerial skill.

## 3.3 Smart beta family classification

We classify the eighty ETFs by smart beta family according to the economic criterion driving their selection rule. The starting classification comes from the data source itself and we have cleaned it in two respects.

First, we harmonised labels that were duplicated by formatting differences. Second, we reclassified four products whose original label did not match their nature: IYY (iShares Dow Jones U.S.) is a broad market ETF rather than a growth product; FDN (First Trust Dow Jones Internet) is a thematic sector fund; KIE (SPDR S&P Insurance) is an insurance sector ETF, notwithstanding its equal-weighted index; and FPX (First Trust US Equity Opportunities) tracks an index of recent IPOs and spin-offs, unrelated to momentum. We group these four in a separate category and exclude them from the family-level analysis, while retaining them in the full-sample analysis.

We also group the smallest families into an *Other factors* category (Quality, Fundamental, Low Volatility and Multi-factor), since disaggregated analysis of groups containing one or two assets would support no statistical inference. Table 1 reports the resulting classification.

**Table 1:** Sample classification by smart beta family.

| **Family** | **N** | **Used in** |
|---|---|---|
| Value | 23 | Full sample and family analysis |
| Growth | 22 | Full sample and family analysis |
| Dividend | 14 | Full sample and family analysis |
| Other factors | 13 | Full sample and family analysis |
| Other (Momentum, ESG, Equal Weight) | 4 | Full sample only |
| Sector / Broad market | 4 | Full sample only |
| **Total** | **80** | |

The family-level analysis therefore covers 72 ETFs across four groups large enough to support conclusions. The appendix lists all eighty products with their names and classification.

## 3.4 Risk factors

Risk factors come from the Kenneth French Data Library. We use the market excess return over the risk-free rate (Mkt-RF), the size (SMB) and value (HML) factors of Fama and French (1993), the profitability (RMW) and investment (CMA) factors added in Fama and French (2015), and the momentum factor (MOM) of Carhart (1997). The risk-free rate (RF) comes from the same source.

We download factors at daily and monthly frequency and express them in decimal form to match the ETF return series.

### 3.5 Return series construction

Daily returns are used in arithmetic form. Monthly series are built from daily log returns, aggregated by summation within each calendar month, which exploits the additivity of logarithmic returns and provides a close approximation to the monthly return on the asset. We drop the final month of the sample as incomplete.

### 3.6 Family portfolios

Alongside the ETF-level analysis, we construct equally weighted portfolios for each smart beta family, together with a portfolio of all eighty products. Portfolio aggregation serves two purposes. It reduces the idiosyncratic noise inherent in individual series, which matters particularly when estimating timing coefficients, notoriously imprecise. And it lets us frame the research question in terms of investment style rather than individual product, while sidestepping the multiple testing problem that arises from examining eighty series at once.

### 3.7 Descriptive statistics

Table 2 reports the main descriptive statistics for the equally weighted family portfolios at monthly frequency, computed on excess returns over the risk-free rate.

**Table 2:** Descriptive statistics of monthly excess returns on the equally weighted portfolios (%). Period 2013:01–2026:02, 158 observations. Sharpe ratios are annualised.

| **Portfolio** | **N** | **Mean** | **Std. dev.** | **Min** | **Max** | **Skewness** | **Sharpe** |
|---|---|---|---|---|---|---|---|
| Value | 23 | 0.594 | 4.925 | −25.471 | 14.861 | −0.964 | 0.418 |
| Growth | 22 | 0.838 | 4.729 | −17.486 | 15.628 | −0.651 | 0.614 |
| Dividend | 14 | 0.519 | 4.031 | −19.698 | 12.855 | −0.863 | 0.446 |
| Other factors | 13 | 0.680 | 4.239 | −20.126 | 13.618 | −0.873 | 0.556 |
| All | 80 | 0.683 | 4.412 | −20.596 | 14.414 | −0.833 | 0.536 |

The statistics reveal differences consistent with the nature of each style. The *Growth* portfolio delivers the highest average return (0.838% monthly) and the highest Sharpe ratio (0.614), reflecting the strong performance of growth stocks over the sample period. *Dividend* displays the most defensive profile, with the lowest average return (0.519%) but also the lowest volatility (4.031%). *Value* shows the highest volatility and the sharpest monthly drawdown (−25.47%, corresponding to the March 2020 collapse).

All portfolios exhibit negative skewness and excess kurtosis, standard features of equity return series, which justify the use of robust standard errors in estimation.

Table 3 reports descriptive statistics for the risk factors.

**Table 3:** Descriptive statistics of the risk factors (% monthly). Period 2013:01–2026:02.

| **Factor** | **Mean** | **Std. dev.** | **Min** | **Max** |
|---|---|---|---|---|
| Mkt-RF | 1.094 | 4.227 | −13.350 | 13.580 |
| SMB | −0.153 | 2.837 | −8.180 | 8.330 |
| HML | −0.053 | 3.485 | −13.830 | 12.860 |
| RMW | 0.236 | 2.064 | −5.810 | 7.190 |
| CMA | −0.056 | 2.243 | −7.080 | 7.730 |
| MOM | 0.233 | 3.674 | −16.210 | 10.000 |
| RF | 0.136 | 0.159 | 0.000 | 0.480 |

One feature of the sample period deserves emphasis: the average market premium is unusually high (1.094% monthly, roughly 13% annually), while size and value premia are on average negative. This should be borne in mind when interpreting our results, since a predominantly bullish period limits how far the conclusions generalise to other market environments.

# 4 Methodology

## 4.1 Overview

It is worth stating at the outset that this thesis addresses two conceptually distinct questions. The first is descriptive: whether ETF returns display the non-linear, time-varying exposure that conventional timing tests are designed to detect. The second is predictive and economic: whether an investor could forecast market states accurately enough to exploit such patterns in real time. Sections 4.2 to 4.7 address the former, Section 4.9 the latter.

Our empirical strategy proceeds through four levels of increasing stringency. We first estimate conventional factor models to determine whether the ETFs generate abnormal returns. We then add market and volatility timing terms to test whether market exposure varies systematically. Third, we examine whether that exposure differs across market states using regime-switching models. Finally, we subject all the evidence to robustness checks and to an out-of-sample economic evaluation.

Throughout, the dependent variable is the excess return of asset $i$ over the risk-free rate, $r_{i,t} = R_{i,t} - R_{f,t}$, and the market factor likewise enters in excess form, $\mathrm{mkt}_t = R_{M,t} - R_{f,t}$. We follow a consistent notational convention: lower-case letters denote excess returns over the risk-free rate and upper-case letters denote total returns, so that $r_{i,t} = R_{i,t} - R_{f,t}$ throughout. The convention applies equally at daily and monthly frequency; only the interpretation of the time subscript changes.

## 4.2 Factor models

We estimate three benchmark specifications. The Fama and French (1993) three-factor model:

$$r_{i,t} = \alpha_i + \beta_i \, \mathrm{mkt}_t + s_i \, SMB_t + h_i \, HML_t + \varepsilon_{i,t}, \tag{1}$$

the Fama and French (2015) five-factor model, which adds profitability and investment:

$$r_{i,t} = \alpha_i + \beta_i \, \mathrm{mkt}_t + s_i \, SMB_t + h_i \, HML_t + r_i \, RMW_t + c_i \, CMA_t + \varepsilon_{i,t}, \tag{2}$$

and an extension of the latter with the Carhart (1997) momentum factor:

$$\begin{aligned} r_{i,t} = \alpha_i &+ \beta_i \, \mathrm{mkt}_t + s_i \, SMB_t + h_i \, HML_t \\ &+ r_i \, RMW_t + c_i \, CMA_t + m_i \, MOM_t + \varepsilon_{i,t}. \end{aligned} \tag{3}$$

The intercept $\alpha_i$ is our measure of abnormal return, or Jensen's alpha: the average return not explained by the fund's exposure to the factors considered. For index-tracking products, a negative alpha reflects the combined effect of management fees and replication costs.

## 4.3 Market timing models

The timing literature builds on the idea that a manager with forecasting ability raises portfolio exposure when anticipating market advances and lowers it otherwise, producing a non-linear relationship between portfolio and market returns.

Treynor and Mazuy (1966) capture this non-linearity through a quadratic term:

$$r_{i,t} = \alpha_i + \beta_i \, \mathrm{mkt}_t + \gamma_i \, \mathrm{mkt}_t^2 + \varepsilon_{i,t}. \tag{4}$$

Marginal market sensitivity is then $\partial r_{i,t}/\partial \mathrm{mkt}_t = \beta_i + 2\gamma_i \mathrm{mkt}_t$, so that $\gamma_i > 0$ implies rising exposure in advancing markets, that is, favourable timing.

Henriksson and Merton (1981) propose an alternative in which exposure takes two values depending on the sign of the market return:

$$r_{i,t} = \alpha_i + \beta_i \mathrm{mkt}_t + \gamma_i \mathrm{mkt}_t^+ + \varepsilon_{i,t}, \qquad \mathrm{mkt}_t^+ = \mathbf{1}\{\mathrm{mkt}_t > 0\} \mathrm{mkt}_t. \tag{5}$$

Here beta equals $\beta_i$ in falling markets and $\beta_i + \gamma_i$ in rising ones, so that $\gamma_i > 0$ again signals favourable timing.

Following Grinblatt and Titman (1994), who caution that performance test results depend on the benchmark employed, we estimate both specifications not only on the market model but also on multifactor bases. Specifically, we use four factor sets: CAPM, Carhart (market, SMB, HML and MOM), the Fama-French five-factor model, and the latter augmented with momentum. The general form is

$$r_{i,t} = \alpha_i + \sum_f \beta_{i,f} f_t + \gamma_i \Psi_t + \varepsilon_{i,t}, \tag{6}$$

where $\Psi_t$ denotes the relevant timing term ($\mathrm{mkt}_t^2$ or $\mathrm{mkt}_t^+$).

## 4.4 Volatility timing and measuring market volatility

Alongside directional timing we test for timing with respect to market volatility, adding a further term:

$$VOL_t = \sigma_{M,t} - \bar{\sigma}_M, \tag{7}$$

where $\sigma_{M,t}$ is market volatility at time $t$ and $\bar{\sigma}_M$ its sample mean. The variable therefore measures how far market volatility stands above or below its typical level. A coefficient $\delta_i > 0$ indicates better-than-usual performance in periods of above-average volatility. We should be precise about what this coefficient captures: since $VOL_t$ enters the return equation additively, $\delta_i$ measures conditional exposure to market volatility rather than the deliberate adjustment of market beta in response to expected volatility, which is what the volatility timing literature traditionally denotes. We therefore refer to $\delta_i$ as a volatility exposure coefficient, while retaining the label volatility timing for the specifications that include it, in line with the empirical literature.

Measuring $\sigma_{M,t}$ warrants specific discussion, since our results prove sensitive to the choice.

At daily frequency we estimate a GARCH(1,1) model on the market excess return,

$$\sigma_{M,t}^2 = \omega + \alpha \varepsilon_{M,t-1}^2 + \beta \sigma_{M,t-1}^2, \tag{8}$$

and take the resulting conditional standard deviation as $\sigma_{M,t}$. With 3,309 observations the model parameters are precisely identified.

At monthly frequency, by contrast, we use realised volatility. Let $r_{t,d}$ denote the daily market excess return on trading day $d$ of month $t$, and let $D_t$ be the number of trading days in that month. The realised monthly volatility for month $t$ is defined as

$$RVOL_t = \left( \sum_{d=1}^{D_t} r_{t,d}^2 \right)^{1/2}, \tag{9}$$

which is the square root of the realised monthly variance $RV_t = \sum_{d=1}^{D_t} r_{t,d}^2$. This definition is appropriate when daily returns are treated as having an approximately zero mean over the month, and follows the convention of constructing realised volatility directly from squared daily returns without demeaning. The volatility term of Section 4.4 is then $VOL_t = RVOL_t - \overline{RVOL}$, where $\overline{RVOL}$ is the sample mean of (9).

Two considerations motivate this choice. The monthly sample contains 158 observations, a relatively short series that makes GARCH parameter estimates potentially imprecise. And realised volatility is a non-parametric estimator that exploits the $D_t$ daily observations within each month, making it considerably more precise (Andersen and Bollerslev, 1998). Our robustness section documents the differences between the two.

## 4.5 The set of estimated models

Combining these specifications yields twenty-three models, estimated at both daily and monthly frequency: three factor models; eight market timing models (Treynor-Mazuy and Henriksson-Merton over four factor bases); the eleven preceding specifications augmented with the volatility timing term; and a market model with volatility timing but no directional timing term, which isolates the volatility effect.

## 4.6 Estimation and inference

All models are estimated by ordinary least squares separately for each ETF and each style portfolio. Since financial return series display heteroskedasticity and autocorrelation, inference relies on Newey and West (1987) standard errors, robust to both. We set the lag length using the automatic rule $L = \lfloor 4(T/100)^{2/9} \rfloor$, which gives eight lags at daily frequency and four at monthly frequency.

## 4.7 Regime-switching models

The models above assume coefficients remain constant throughout the sample. Following Turtle and Zhang (2012), we relax this assumption by allowing alpha, the factor loadings and the timing coefficients to take different values depending on the state of the market.

We identify states using a Markov regime-switching model (Hamilton, 1989) estimated on the market excess return. We assume $\mathrm{mkt}_t$ is normally distributed with mean and variance depending on a latent state variable $S_t \in \{1, 2\}$:

$$\mathrm{mkt}_t = \mu_{S_t} + \varepsilon_t, \qquad \varepsilon_t \sim N(0, \sigma^2_{S_t}), \tag{10}$$

governed by a Markov chain with transition matrix

$$\mathbf{P} = \begin{pmatrix} p_{11} & 1 - p_{22} \\ 1 - p_{11} & p_{22} \end{pmatrix}, \qquad p_{kk} = \Pr(S_t = k \mid S_{t-1} = k), \tag{11}$$

so that the expected duration of each regime is $1/(1 - p_{kk})$. Parameters are estimated jointly by maximum likelihood using the Hamilton filter.

Following Turtle and Zhang (2012), we label as the bear regime the state with the lower estimated mean market excess return, and as bull the other. Each period is assigned to a state according to its smoothed probability. We estimate the model once on the market and apply the resulting state calendar to all eighty ETFs, so that bull and bear carry the same meaning throughout the sample.

We then estimate eight specifications in which every coefficient switches by regime through interaction with state dummies, where $S$ denotes the bull or bear state:

$$r_{i,t} = \sum_{S \in \{\text{bull},\text{bear}\}} D_t^S \left[ \alpha_i^S + \sum_f \beta_{i,f}^S f_t + \gamma_i^S \, \Psi_t + \delta_i^S \, VOL_t \right] + \varepsilon_{i,t}. \tag{12}$$

The variables $D_t^S$ are dummies identifying the regime prevailing in each period. Specifically, $D_t^{\text{bull}}$ equals one if month $t$ is classified as bull and zero otherwise, and $D_t^{\text{bear}}$ works symmetrically. Since each period belongs to one and only one state, the two are mutually exclusive and always sum to one.

The mechanics are straightforward: in each observation only one bracket is active, while the other is multiplied by zero and drops out. Bull-state coefficients are therefore estimated from bull-state observations and bear-state coefficients from the remainder, all within a single regression.

Estimating jointly rather than running two separate regressions on each subsample provides a convenient way to obtain the covariance matrix required for formal cross-regime comparisons. Joint estimation delivers the covariance between the two sets of coefficients, which enters the Wald test on the difference $\gamma_i^{\text{bear}} - \gamma_i^{\text{bull}}$; separate subsample regressions would require additional assumptions or a further step to recover it.

Note finally that the model includes no separate intercept. Since the two dummies sum to one in every period, adding a constant would create perfect collinearity and render the model inestimable. The intercepts $\alpha_i^{\text{bull}}$ and $\alpha_i^{\text{bear}}$ serve that role, one in each regime.

## 4.8 Multiple testing and the block bootstrap

Estimating twenty-three models on eighty assets raises a multiple testing problem that conventional inference ignores. At the 5% significance level we would expect roughly four spurious rejections per model by chance alone. As Lo and MacKinlay (1990) warn, and as Kosowski *et al.* (2006) and Fama and French (2010) show in the fund evaluation context, a substantial share of apparent skill may simply reflect luck.

To address this we implement a block bootstrap with 1,000 replications. The design is as follows. For each asset we estimate the model under the null of no abnormal return and no timing, imposing $\alpha_i = 0$ and excluding the timing terms, and retain the fitted values and residuals. We then generate artificial series by adding to the fitted values the residuals resampled in blocks of six months, which preserves the autocorrelation structure and volatility clustering that independent resampling would destroy. On each artificial series we re-estimate the full model and record the $t$-statistics of the coefficients of interest.

The test compares the observed $t$-statistics against percentiles of the bootstrap distribution of the maximum and minimum across the eighty assets. We should be explicit about the scope of this correction: it controls the family-wise error rate across the eighty ETFs within each model and coefficient separately, not jointly across all twenty-three models and every coefficient simultaneously. A correction over that wider family would be more stringent still. Appendix A sets out the rationale and mechanics of this procedure in detail.

## 4.9 Out-of-sample evaluation

The final block of the analysis tests whether the evidence obtained can be economically exploited. The design strictly respects the information available at each point in time, avoiding look-ahead bias.

Before describing it, we should explain why our trading signals differ from the coefficients estimated above. The timing coefficient $\gamma_i$ is a descriptive measure: it characterises how a fund's market exposure varied over the sample, but says nothing about future market movements. Knowing that a given ETF has $\gamma_i < 0$ tells us something about that product, not about whether to be invested next month. It cannot therefore serve as a trading rule.

The only genuinely forward-looking quantity in our analysis is the probability of being in a given market regime, obtained from the regime-switching model, since it refers to the state of the market rather than to a fund's past behaviour. Our economic evaluation is accordingly built on that signal and, as an alternative, on lagged realised volatility.

The question tested here is thus complementary to the preceding one. Earlier sections ask whether funds *themselves* display systematically varying market exposure; this section asks whether an investor could *generate* such variation by anticipating the market state and adjusting position accordingly. Both address the same underlying hypothesis, the existence of exploitable timing, from different angles.

The evaluation period covers the final 98 months of the sample (January 2018 to February 2026), reserving the first 60 observations as an initial estimation window. In each month $t$ of the evaluation period, parameters are estimated using information up to $t-1$ only, with the window expanding progressively.

We consider two investment signals. The first is the filtered probability of being in the bear regime, obtained by re-estimating the regime-switching model at each point with information up to $t-1$. We stress that we use filtered rather than smoothed probabilities: the latter incorporate information from the whole sample, including periods after the date being evaluated, and are therefore valid for description but not for prediction. The second signal is lagged realised volatility, compared with its recursively computed historical mean.

Both signals generate binary investment rules: the position is either fully invested in the portfolio or fully moved to the risk-free asset, with no intermediate positions, leverage or short selling. The specific rules are as follows.

The regime strategy stays invested when $p_t^{\text{bear}} \leq 0.5$ and moves to the risk-free asset when $p_t^{\text{bear}} > 0.5$, where $p_t^{\text{bear}}$ is the filtered bear-regime probability obtained with information up to $t-1$.

The volatility strategy stays invested when $RVOL_{t-1} \leq \overline{RVOL}_{t-1}$ and moves to the risk-free asset when $RVOL_{t-1} > \overline{RVOL}_{t-1}$, where $RVOL_{t-1}$ is the previous month's realised volatility as defined in (9) and $\overline{RVOL}_{t-1}$ its mean computed on an expanding basis over all preceding months, so that this reference also uses past information only.

The difference between the two is not merely one of threshold. The first rests on a quantity *estimated* through an econometric model, the probability of an unobservable latent state; the second on a directly *observable* statistic, with no intervening model that could be misspecified. Testing both allows us to distinguish whether any failure stems from limitations of the regime model or from the absence of predictive content in volatility information, however measured.

The resulting strategies are compared against a passive buy-and-hold position using annualised Sharpe ratios and cumulative returns.

Finally, we incorporate transaction costs of ten basis points per unit of turnover. This correction is essential: timing strategies involve substantial turnover, and omitting costs systematically overstates attainable returns.

Costs are applied by deducting them from the return of the period in which trading occurs. Denoting by $w_t \in \{0, 1\}$ the position taken, the net total return is

$$R_t^{\text{net}} = w_t \, R_t + (1 - w_t) \, R_{f,t} - |w_t - w_{t-1}| \cdot c, \tag{13}$$

with $c = 0.001$, where $R_t$ is the total return on the portfolio, consistent with the notational convention set out in Section 4.1. Sharpe ratios are computed on the corresponding excess return, $R_t^{\text{net}} - R_{f,t}$. The term $|w_t - w_{t-1}|$ equals one in months when the strategy changes position, whether entering or exiting the market, and zero when it does not, so costs are incurred only when trading. All return and Sharpe ratio figures reported in the results section are net of these costs.

Two clarifications are in order. Buy-and-hold incurs no cost under this formulation, since it takes a single position at the outset and never changes it. Nor do we charge for the implicit rebalancing involved in maintaining equal weights within the portfolios. We should not claim that this omission is neutral across strategies: timing strategies spend part of the period in cash, and selected portfolios may involve different constituent turnover, so the omitted charge need not fall equally on all alternatives. We judge it unlikely to be decisive for the comparisons drawn here, but the assumption is one of convenience rather than one we can justify as innocuous.

The ten basis point figure itself requires justification. It approximates the combined bid-ask spread and market impact for the more heavily traded ETFs in our sample, most of which track broad indices from established providers and typically show spreads below five basis points. We should note, however, that this is a conservative estimate: thinly traded products, concentrated in the *Other factors* category and among pure-style funds, may exhibit substantially wider spreads and less market depth, which would raise the impact of orders. To the extent that actual costs exceed our assumption, timing strategies would perform even worse than documented here, so the choice does not compromise our conclusions but reinforces them. We use a single moderate figure rather than product-specific estimates to avoid introducing further assumptions about liquidity data unavailable in our sample.

# 5 Results

We present the empirical evidence in order of increasing stringency. We begin with factor models and timing tests estimated under conventional inference, then turn to the family-level and regime-switching results, and finally subject all of it to robustness checks, multiple testing correction and an out-of-sample economic evaluation.

## 5.1 Abnormal returns: factor models

Table 4 reports Jensen's alpha estimates for the equally weighted style portfolios at monthly frequency.

**Table 4:** Alpha of the equally weighted family portfolios (% monthly). HAC standard errors. Significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

| Model | Value | Growth | Dividend | Other factors | All |
|---|---|---|---|---|---|
| FF3 | −0.426*** | −0.254** | −0.376*** | −0.291*** | −0.331*** |
| FF5 | −0.432*** | −0.236** | −0.408*** | −0.308*** | −0.333*** |
| FF5+MOM | −0.430*** | −0.286*** | −0.431*** | −0.331*** | −0.359*** |

The result is unambiguous: every portfolio delivers a negative and statistically significant alpha across all three specifications. Magnitudes range from −0.24% to −0.43% monthly. Annualising by simple multiplication ($12\alpha$), this corresponds to a shortfall of between 2.8% and 5.2% per year relative to what the factor models predict; compounding would give marginally smaller figures, though the difference is immaterial at these magnitudes.

A clear ordering emerges across styles. *Value* and *Dividend* display the most negative alphas, while *Growth* is the least unfavourable. This admits a direct economic reading: the HML and RMW factors in the five-factor model capture precisely the exposure characteristic of value and dividend products, so these add nothing beyond that exposure and their alpha reflects management and replication costs alone. Growth products, which depart further from the factor combination considered, show smaller and less systematic alphas.

Individual estimates reinforce this conclusion. Under the five-factor model, all ETFs in the Value (23 of 23) and Dividend (14 of 14) families display negative alphas significant at the 5% level, against 9 of 22 in *Growth* and 9 of 13 in *Other factors*. Not a single ETF shows a positive significant alpha in any specification, family or frequency.

## 5.2 Market timing

Table 5 reports market timing coefficients for the style portfolios.

**Table 5:** Market timing coefficient ($\gamma$) of the equally weighted family portfolios, monthly frequency. HAC standard errors.

| Model | Value | Growth | Dividend | Other factors | All |
|---|---|---|---|---|---|
| TM-CAPM | −1.00 | −0.63* | −0.93 | −0.92 | −0.83 |
| HM-CAPM | −0.19 | −0.14* | −0.22 | −0.20 | −0.18 |
| TM-FF5+MOM | −0.35 | −0.56* | −0.44 | −0.56* | −0.45 |
| HM-FF5+MOM | −0.09 | −0.13** | −0.14** | −0.14** | −0.11* |

Timing coefficients are negative without exception across portfolios and specifications. A negative $\gamma$ means effective market exposure falls in advancing markets and rises in declining ones, an unfavourable timing pattern. Under the Henriksson-Merton specification with five factors and momentum, this pattern reaches statistical significance in the *Growth*, *Dividend* and *Other factors* portfolios, as well as in the full portfolio.

We stress that this cannot be read as managerial ability, for the reasons set out in the data section: these products track rules-based indices with no discretionary exposure decisions. The negative coefficient instead reflects the mechanical convexity that index construction induces in the relationship between portfolio and market.

## 5.3 Volatility timing

Adding the volatility timing term produces two noteworthy results. First, the coefficient $\delta$ is negative across all portfolios, reaching significance in *Value* (−0.258) and in the full portfolio (−0.154). At the individual level, under the market model with volatility timing, 12 of the 23 Value ETFs display a significant negative $\delta$, with no positive cases. In months when realised market volatility exceeds its average level, these products underperform what their factor exposure would predict.

Second, and of greater methodological interest, including the volatility term absorbs much of the apparent directional timing. Under TM-FF5+MOM, ten of the twenty-two Growth ETFs displayed a significant negative $\gamma$; once $VOL_t$ enters the model, that number falls to zero. This suggests that a substantial part of what conventional market timing tests identify as directional timing is in fact sensitivity to market volatility. Section 5.5.2 decomposes this effect and shows that coefficient reassignment is compounded by an increase in estimator variance arising from the correlation between the two terms.

## 5.4 Regime-switching results

### 5.4.1 Identifying the states

The regime-switching model estimated on market excess returns identifies two clearly distinct states. At monthly frequency, the bear regime shows a mean of 0.42% and volatility of 5.60%, with an expected duration of roughly eight months, against 1.68% mean and 2.27% volatility in the bull regime. The sample splits into 68 bear months and 90 bull months.

Periods are assigned to states using smoothed probabilities, which condition on the entire sample. These are therefore retrospectively classified states, appropriate for ex-post description but not usable in real time; Section 5.7 employs filtered probabilities instead, precisely because the out-of-sample exercise requires a signal available at the moment of decision.

Figure 1 validates this identification. Periods classified as bear coincide with the corrections actually observed: the 2015–2016 slowdown, the fourth quarter of 2018, the pandemic collapse of 2020, the 2022 bear market and the 2025 correction.

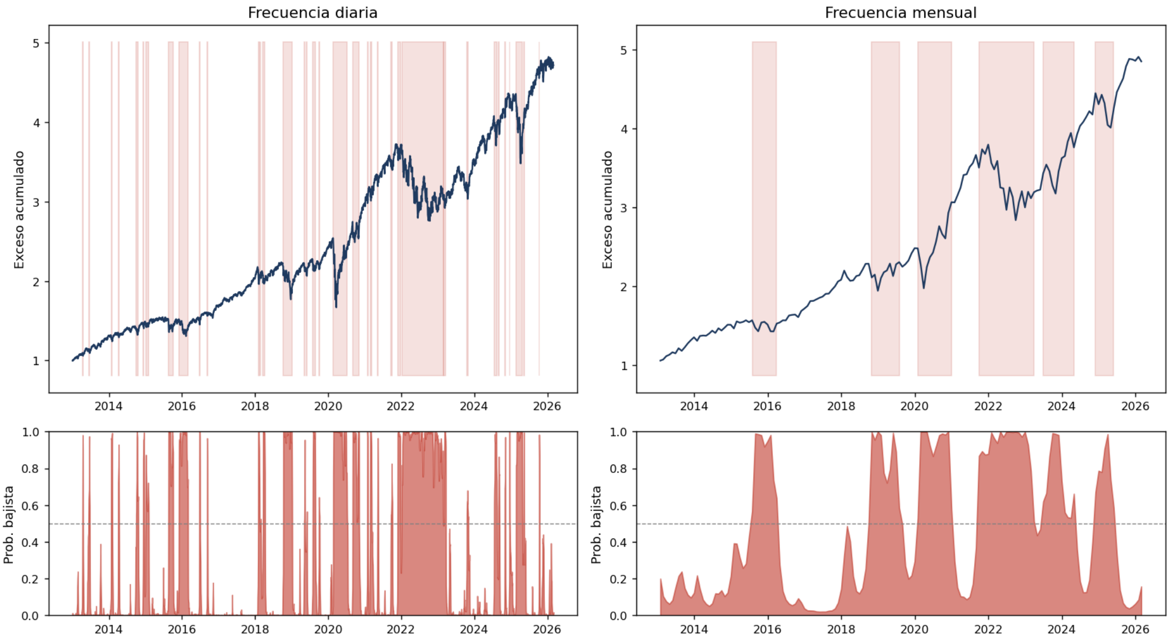


**Figure 1:** Regime identification. The upper panel plots cumulative market excess returns, $\prod_t (1+\text{mkt}_t)$ starting from unity, with periods classified as bear shaded; the lower panel shows the smoothed probability of being in that state. Daily frequency (left) and monthly frequency (right).

One difference between frequencies deserves note. In daily data the states separate mainly on volatility, producing short and frequent episodes; in monthly data the separation also reflects an appreciable difference in means, generating longer and better-defined phases. This justifies basing the regime analysis primarily on monthly frequency.

#### 5.4.2 State-dependent timing

Table 6 reports timing coefficients estimated separately within each regime.

**Table 6:** Market timing coefficient by regime, equally weighted family portfolios, monthly frequency. The final column reports the $p$-value of the test of equality across states. HAC standard errors.

| **Portfolio** | $\gamma^{\text{bull}}$ | $\gamma^{\text{bear}}$ | **$p$-value diff.** |
|---|---|---|---|
| *Panel A: Treynor-Mazuy with volatility timing* | | | |
| Value | −4.41 | +1.04 | 0.069 |
| Growth | +3.74* | −0.44 | **0.046** |
| Dividend | −6.00*** | +0.36 | **0.008** |
| Other factors | −2.68 | +0.22 | 0.119 |
| All | −1.38 | +0.27 | 0.322 |
| *Panel B: Henriksson-Merton with volatility timing* | | | |
| Value | −0.48 | +0.22 | 0.073 |
| Growth | +0.53*** | −0.17 | **0.004** |
| Dividend | −0.64** | +0.02 | **0.036** |
| Other factors | −0.28 | +0.00 | 0.277 |
| All | −0.11 | +0.01 | 0.620 |

The most striking result is that Growth is the only family whose timing behaviour differs systematically across states, with a positive significant coefficient in the bull regime and a negative one in the bear regime, the difference being statistically significant. At the individual level, fifteen of the twenty-two Growth ETFs show significant cross-regime differences, against none in the Value family. The Dividend portfolio also shows significant differences, though with negative coefficients in both regimes and larger magnitude in the bull state.

This pattern is consistent with the construction of growth indices, which concentrate exposure in high-beta securities with implicit momentum: exposure amplifies during advances and turns unfavourable during declines. Value, by contrast, behaves stably across states.

We should be clear, however, that this does not constitute evidence of timing ability. A genuine market timer would display a positive $\gamma$ in *both* states, raising exposure ahead of advances and protecting during declines. What Growth displays is procyclicality: amplification when markets rise and no protection when they fall, which is close to the opposite of what an investor would want.

## 5.5 Cross-sectional distribution of the coefficients

The preceding sections have reported style portfolio results and some aggregate counts. Given the number of specifications estimated, it is useful to summarise the distribution of coefficients across the eighty ETFs for all twenty-three models, following the format of Fung, Xu and Yau (2002). Tables 7, 8 and 9 report, for each model, the mean, quartiles and median of the relevant coefficient, together with the number of positive and negative estimates and, in parentheses, how many reach significance at the 5% level. Appendix B extends this information for the main specifications.

Reporting medians and quartiles matters because the mean can be pulled by extreme observations, particularly for the Treynor-Mazuy coefficient, whose scale amplifies outliers.

**Table 7:** Cross-sectional distribution of alpha (% monthly). Cross-sectional distribution of the coefficient across the eighty ETFs, monthly frequency. $N^+$ and $N^-$ report the number of positive and negative coefficients; in parentheses, how many are significant at the 5% level.

| Model | Mean | Q1 | Median | Q3 | $N^+(N^{+*})$ | $N^-(N^{-*})$ |
|---|---|---|---|---|---|---|
| FF3 | −0.331 | −0.404 | −0.335 | −0.251 | 0 (0) | 80 (56) |
| FF5 | −0.333 | −0.415 | −0.347 | −0.249 | 0 (0) | 80 (59) |
| FF5+MOM | −0.359 | −0.434 | −0.370 | −0.309 | 0 (0) | 80 (65) |
| TM-CAPM | −0.249 | −0.383 | −0.259 | −0.125 | 11 (0) | 69 (11) |
| HM-CAPM | −0.108 | −0.246 | −0.071 | 0.028 | 24 (0) | 56 (2) |
| TM-Carhart | −0.259 | −0.348 | −0.289 | −0.191 | 8 (0) | 72 (38) |
| HM-Carhart | −0.152 | −0.259 | −0.188 | −0.028 | 15 (0) | 65 (2) |
| TM-FF5 | −0.234 | −0.338 | −0.242 | −0.121 | 8 (0) | 72 (30) |
| HM-FF5 | −0.120 | −0.235 | −0.145 | 0.004 | 21 (0) | 59 (5) |
| TM-FF5+MOM | −0.273 | −0.353 | −0.294 | −0.205 | 3 (0) | 77 (42) |
| HM-FF5+MOM | −0.167 | −0.278 | −0.196 | −0.057 | 14 (0) | 66 (3) |
| FF3+VOL | −0.306 | −0.379 | −0.320 | −0.232 | 0 (0) | 80 (55) |
| FF5+VOL | −0.310 | −0.396 | −0.323 | −0.237 | 0 (0) | 80 (57) |
| FF5+MOM+VOL | −0.334 | −0.410 | −0.349 | −0.286 | 0 (0) | 80 (64) |
| TM-CAPM+VOL | −0.363 | −0.500 | −0.371 | −0.190 | 7 (0) | 73 (15) |
| HM-CAPM+VOL | −0.301 | −0.481 | −0.311 | −0.099 | 16 (0) | 64 (4) |
| TM-Carhart+VOL | −0.337 | −0.416 | −0.380 | −0.249 | 0 (0) | 80 (43) |
| HM-Carhart+VOL | −0.269 | −0.382 | −0.287 | −0.145 | 11 (0) | 69 (7) |
| TM-FF5+VOL | −0.302 | −0.408 | −0.331 | −0.175 | 0 (0) | 80 (38) |
| HM-FF5+VOL | −0.229 | −0.350 | −0.247 | −0.103 | 17 (0) | 63 (10) |
| TM-FF5+MOM+VOL | −0.346 | −0.431 | −0.387 | −0.264 | 1 (0) | 79 (48) |
| HM-FF5+MOM+VOL | −0.276 | −0.391 | −0.303 | −0.130 | 11 (0) | 69 (11) |
| CAPM+VOL | −0.362 | −0.500 | −0.346 | −0.245 | 1 (0) | 79 (38) |

**Table 8:** Cross-sectional distribution of the market timing coefficient ($\gamma$). Cross-sectional distribution of the coefficient across the eighty ETFs, monthly frequency. $N^+$ and $N^-$ report the number of positive and negative coefficients; in parentheses, how many are significant at the 5% level.

| Model | Mean | Q1 | Median | Q3 | $N^+(N^{+*})$ | $N^-(N^{-*})$ |
|---|---|---|---|---|---|---|
| TM-CAPM | −0.828 | −1.212 | −0.776 | −0.551 | 3 (0) | 77 (7) |
| HM-CAPM | −0.177 | −0.230 | −0.179 | −0.132 | 2 (0) | 78 (7) |
| TM-Carhart | −0.487 | −0.754 | −0.518 | −0.153 | 10 (0) | 70 (16) |
| HM-Carhart | −0.119 | −0.169 | −0.116 | −0.068 | 3 (0) | 77 (19) |
| TM-FF5 | −0.539 | −0.787 | −0.543 | −0.285 | 4 (0) | 76 (23) |
| HM-FF5 | −0.129 | −0.174 | −0.122 | −0.083 | 2 (0) | 78 (24) |
| TM-FF5+MOM | −0.449 | −0.720 | −0.433 | −0.183 | 6 (0) | 74 (15) |
| HM-FF5+MOM | −0.114 | −0.164 | −0.109 | −0.071 | 6 (0) | 74 (20) |
| TM-CAPM+VOL | 0.005 | −0.295 | 0.005 | 0.323 | 40 (0) | 40 (0) |
| HM-CAPM+VOL | −0.038 | −0.091 | −0.044 | 0.015 | 27 (0) | 53 (3) |
| TM-Carhart+VOL | 0.063 | −0.197 | 0.145 | 0.311 | 48 (0) | 32 (0) |
| HM-Carhart+VOL | −0.035 | −0.084 | −0.032 | 0.013 | 24 (0) | 56 (1) |
| TM-FF5+VOL | −0.050 | −0.335 | 0.011 | 0.276 | 41 (0) | 39 (1) |
| HM-FF5+VOL | −0.051 | −0.099 | −0.050 | 0.005 | 24 (0) | 56 (5) |
| TM-FF5+MOM+VOL | 0.070 | −0.180 | 0.152 | 0.309 | 48 (0) | 32 (0) |
| HM-FF5+MOM+VOL | −0.036 | −0.092 | −0.036 | 0.018 | 23 (0) | 57 (3) |

**Table 9:** Cross-sectional distribution of the volatility timing coefficient ($\delta$). Cross-sectional distribution of the coefficient across the eighty ETFs, monthly frequency. $N^+$ and $N^-$ report the number of positive and negative coefficients; in parentheses, how many are significant at the 5% level.

| Model | Mean | Q1 | Median | Q3 | $N^+(N^{+*})$ | $N^-(N^{-*})$ |
|---|---|---|---|---|---|---|
| FF3+VOL | −0.099 | −0.144 | −0.092 | −0.057 | 1 (0) | 79 (41) |
| FF5+VOL | −0.101 | −0.123 | −0.103 | −0.073 | 0 (0) | 80 (52) |
| FF5+MOM+VOL | −0.097 | −0.117 | −0.098 | −0.070 | 0 (0) | 80 (52) |
| TM-CAPM+VOL | −0.154 | −0.205 | −0.153 | −0.046 | 0 (0) | 80 (5) |
| HM-CAPM+VOL | −0.143 | −0.182 | −0.132 | −0.047 | 0 (0) | 80 (8) |
| TM-Carhart+VOL | −0.102 | −0.141 | −0.102 | −0.073 | 4 (0) | 76 (14) |
| HM-Carhart+VOL | −0.088 | −0.131 | −0.089 | −0.048 | 10 (0) | 70 (22) |
| TM-FF5+VOL | −0.097 | −0.122 | −0.098 | −0.067 | 2 (0) | 78 (22) |
| HM-FF5+VOL | −0.086 | −0.109 | −0.081 | −0.058 | 0 (0) | 80 (30) |
| TM-FF5+MOM+VOL | −0.102 | −0.124 | −0.098 | −0.071 | 0 (0) | 80 (21) |
| HM-FF5+MOM+VOL | −0.086 | −0.109 | −0.081 | −0.058 | 0 (0) | 80 (30) |
| CAPM+VOL | −0.154 | −0.191 | −0.142 | −0.078 | 0 (0) | 80 (27) |

Three regularities stand out. First, no model produces a single positive significant alpha, while under the pure factor specifications all eighty products show negative alphas, with between fifty-six and sixty-five significant estimates. The entire distribution lies in negative territory: even the third quartile falls below zero in virtually every model.

Second, the timing coefficient is distinctly negative in specifications without the volatility term, with medians between $-0.11$ and $-0.83$ and between seventy and seventy-eight negative estimates out of eighty. Adding the volatility term shifts the distribution towards zero, with estimates splitting roughly evenly between signs and significant cases all but disappearing, consistent with sections 5.3 and 5.5.2.

Third, the volatility timing coefficient is uniformly negative, with medians near $-0.10$ and almost no positive estimates, indicating that unfavourable sensitivity to volatility is a general feature of the sample.

## 5.6 Robustness

### 5.6.1 Measuring volatility

Comparing the GARCH model against realised volatility reveals considerable sensitivity to the measure employed. The two agree on the average level of monthly volatility (4.26% against 4.24%), but their correlation is only 0.468, indicating appreciably different time paths.

The consequences for the coefficients matter. Under the GARCH specification the Growth portfolio's timing coefficient is significant; using realised volatility it is not. Conversely, volatility timing coefficients barely reach significance under the GARCH measure but do so under realised volatility. For the reasons given in the methodology we adopt realised volatility as our main specification, while acknowledging explicitly that conclusions about timing are sensitive to how volatility is measured.

### 5.6.2 Separating non-linear market exposure from volatility exposure

The fall in the number of significant timing coefficients when the volatility term is added, documented in section 5.3, admits two explanations worth separating. The effect may be reassigned to the volatility term, or the correlation between the two regressors may inflate estimator variance and make individual significance harder to attain.

The two regressors are indeed related: the quadratic term $\mathrm{mkt}_t^2$ takes its largest values when market returns are extreme in absolute terms, which coincides with periods of high volatility. Their sample correlation is 0.568 at monthly frequency.

This correlation does not, however, compromise identification. Table 10 reports variance inflation factors, all well below conventionally problematic thresholds and comparable to those of routinely used factors such as HML or CMA.

**Table 10:** Variance inflation factors for the TM-FF5+MOM+VOL model, monthly frequency.

| Mkt-RF | SMB | HML | RMW | CMA | MOM | mkt$^2$ | $VOL$ |
|---|---|---|---|---|---|---|---|
| 1.53 | 1.64 | 2.44 | 1.31 | 2.05 | 1.44 | 1.76 | 2.13 |

To discriminate between the two explanations we decompose the effect across the twenty-two Growth ETFs, where it is most pronounced. Adding the volatility term reduces the average timing coefficient from $-0.565$ to $-0.157$, a contraction of 72%; in absolute terms the average magnitude falls by 43%. At the same time, the average standard error rises by 62%.

Both mechanisms are therefore at work, and together they point to an identification difficulty rather than a purely statistical nuisance. Part of the effect is genuinely reassigned to the volatility term, as the coefficient contraction indicates, and part of the loss of significance reflects greater estimator variance arising from the correlation between regressors. The relevant conclusion is that the two regressors are capturing the same underlying phenomenon, namely portfolio behaviour during turbulent market conditions, and that the data do not allow non-linear market exposure to be cleanly separated from volatility exposure. This is an empirical identification problem, not merely a loss of precision. This anticipates the result of the following section, where neither coefficient survives multiple testing correction.

### 5.6.3 Temporal stability

Splitting the sample before and after 2020 yields an equally instructive diagnosis. The negative alpha remains stable across both subperiods, across all families and at comparable magnitudes.

Timing coefficients, by contrast, show no stability whatsoever. The Value portfolio coefficient moves from $-0.46$ in the first subperiod to $+0.48$ in the second, reversing sign. Dividend shows a significant $-1.24$ before 2020 that disappears thereafter. No timing coefficient is significant in both subperiods simultaneously.

## 5.7 Multiple testing correction

The results above rest on conventional inference, which ignores that eighty hypotheses are being tested simultaneously. Table 11 compares the number of significant coefficients under both criteria.

**Table 11:** Number of ETFs (of 80) with coefficients significant at the 5% level under conventional inference and under a block bootstrap with 1,000 replications. Monthly frequency.

| **Model** | **Par.** | **Conventional** | **Threshold** | **Bootstrap** |
|---|---|---|---|---|
| FF5+MOM | $\alpha$ | 65 | −3.30 | **20** |
| TM-FF5+MOM | $\alpha$ | 42 | −3.29 | 0 |
| TM-FF5+MOM | $\gamma$ | 15 | −3.89 | **0** |
| HM-FF5+MOM | $\gamma$ | 20 | −3.25 | **1** |
| TM-FF5+MOM+VOL | $\alpha$ | 48 | −3.25 | 1 |
| TM-FF5+MOM+VOL | $\delta$ | 21 | −3.69 | **0** |
| CAPM+VOL | $\delta$ | 27 | −3.93 | **1** |

The correction changes the conclusions substantially. The critical threshold moves from the conventional −1.96 to values near −3.3, reflecting that when eighty assets are examined at once, some will reach considerable statistics by chance alone.

Figure 2 presents this graphically. Each panel superimposes two histograms. The grey one is the bootstrap distribution: each of its values is the most extreme $t$-statistic reached by any of the eighty ETFs in one of the thousand replications where, by construction, no effect exists. It concentrates around −2, meaning that in a universe with neither abnormal returns nor timing, it is normal for some asset to land near the conventional threshold purely by chance. The red histogram shows the $t$-statistics actually observed in the real sample.

The two vertical lines mark the decision criteria: the dotted line is the conventional −1.96 threshold and the dashed line the bootstrap threshold, at −3.30 for alpha and −3.89 for the timing coefficient. The region between them is particularly telling, since it contains coefficients that conventional inference would declare significant but that are indistinguishable from chance. In the left panel, for alpha, a substantial share of red bars lies to the left of the dashed line and therefore survives the strict criterion. In the right panel, for the timing coefficient, almost no red bar reaches that threshold, even though many clear the conventional one.

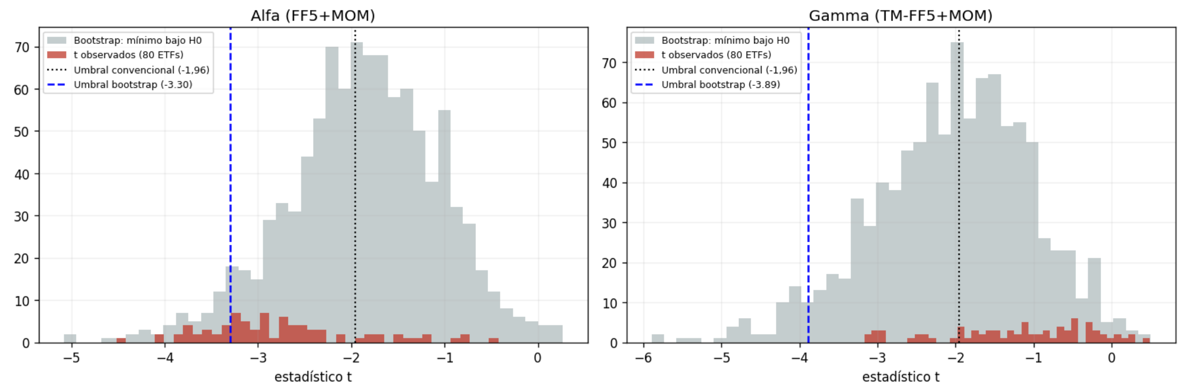


**Figure 2:** Bootstrap distribution of the minimum $t$-statistic under the null (grey) against the $t$-statistics observed across the eighty ETFs (red). Each value in the grey histogram is the most extreme statistic reached by any asset in a replication with no real effect. The dotted line marks the conventional threshold ($-1.96$) and the dashed line the bootstrap threshold. Only coefficients to the left of the latter can be considered significant.

Few results survive this correction, but those that do matter. The negative alpha holds: twenty of the eighty ETFs remain significant under the strict criterion, concentrated in Value (ten of twenty-three) and Dividend (four of fourteen). By contrast, timing evidence all but vanishes: the fifteen to twenty significant $\gamma$ coefficients under conventional inference fall to zero or one, and the same holds for volatility timing.

We should be precise about what this means. Losing significance under the corrected criterion does not establish that the underlying coefficients are zero. It means that, at the chosen family-wise error rate, the data no longer provide sufficient evidence to distinguish these estimates from sampling variation. The instability found in the robustness checks points in the same direction: the absence of a stable pattern across specifications and subperiods is what one would expect if there is no underlying phenomenon of the kind the tests are designed to detect.

## 5.8 Out-of-sample evaluation

Our final test asks whether the evidence obtained can be economically exploited. As set out in section 4.9, the strategies are built on regime probabilities and lagged volatility rather than on the estimated timing coefficients, given the descriptive nature of the latter. Table 12 reports results over the 98 months of the evaluation period.

**Table 12:** Out-of-sample evaluation (2018:01–2026:02, 98 months). Annualised Sharpe ratios and cumulative returns. Strategies are net of transaction costs of 10 basis points per unit of turnover.

| | **Buy and hold** | | **Regime strategy** | | **Volatility strategy** | |
|---|---|---|---|---|---|---|
| **Portfolio** | Sharpe | Cum. | Sharpe | Cum. | Sharpe | Cum. |
| Value | **0.256** | **57.1%** | 0.002 | 17.7% | −0.186 | −0.6% |
| Growth | **0.446** | **110.7%** | −0.043 | 12.3% | 0.047 | 21.9% |
| Dividend | **0.237** | **50.5%** | −0.014 | 17.5% | −0.119 | 8.3% |
| Other factors | **0.356** | **78.1%** | 0.028 | 21.0% | −0.038 | 14.9% |
| All | **0.349** | **78.3%** | −0.010 | 17.3% | −0.073 | 11.4% |

No strategy outperforms the passive benchmark, in any portfolio and under either signal. The magnitude of the gap is considerable: in the Growth portfolio, the regime-based strategy cuts cumulative returns from 110.7% to 12.3%, destroying roughly 89% of what a passive position would have delivered. Sharpe ratios fall to values near zero or negative throughout.

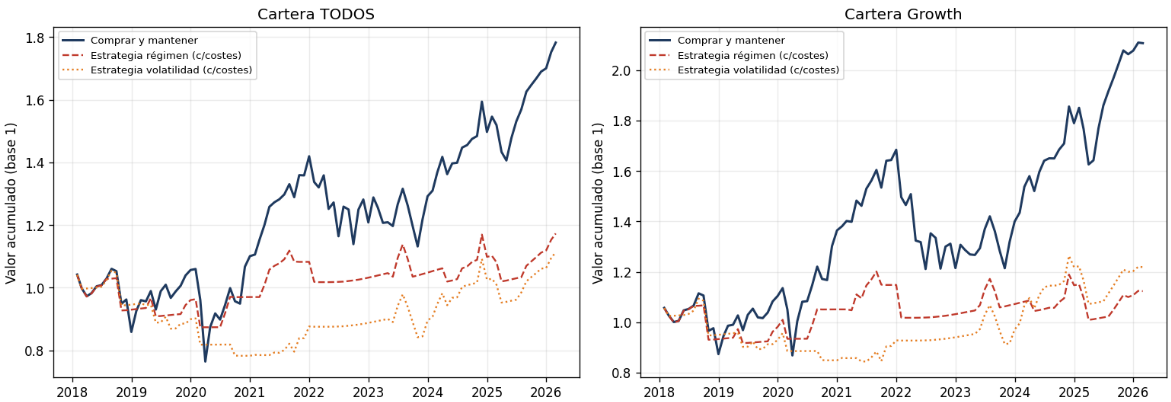


**Figure 3:** Cumulative value of the timing strategies against the passive position over the out-of-sample evaluation period, net of transaction costs.

The explanation lies in how often the signals recommend exiting the market: the regime signal holds the portfolio out for 53 of the 98 months evaluated, and the volatility signal for 47. In a predominantly rising market, staying out for more than half the time is extremely costly, and the signals do not identify unfavourable phases precisely enough to compensate. Transaction costs merely worsen an outcome that is already poor without them.

### 5.8.1 A more elaborate decision rule

One might object that the strategies above rely on an elementary binary rule that considers only the most likely state and discards the rest of the available information. A month with a bear probability of 0.51 and one with 0.95 receive identical treatment, and the magnitude of expected returns in each state plays no role.

We therefore test a more elaborate rule, close to that of Turtle and Zhang (2012), computing expected returns as a probability-weighted average of the two scenarios:

$$E[r_t] = p_t^{\text{bull}}\left(\hat{\alpha}^{\text{bull}} + \hat{\beta}^{\text{bull}}\hat{\mu}^{\text{bull}}\right) + p_t^{\text{bear}}\left(\hat{\alpha}^{\text{bear}} + \hat{\beta}^{\text{bear}}\hat{\mu}^{\text{bear}}\right), \tag{14}$$

where $p_t^S$ are filtered probabilities, $\hat{\mu}^S$ the market means estimated in each regime, and $\hat{\alpha}^S, \hat{\beta}^S$ the portfolio coefficients by state. All components are re-estimated in a rolling window with information up to $t-1$. The position stays invested when $E[r_t] > 0$, that is, when the portfolio is expected to beat the risk-free asset.

**Table 13:** Strategy based on expected returns. Annualised Sharpe ratios and cumulative returns net of costs.

| | **Buy and hold** | | **Expected return** | |
|---|---|---|---|---|
| **Portfolio** | Sharpe | Cum. | Sharpe | Cum. |
| Value | **0.256** | **57.1%** | −0.035 | 10.2% |
| Growth | **0.446** | **110.7%** | −0.019 | 6.4% |
| Dividend | **0.237** | **50.5%** | 0.073 | 24.1% |
| Other factors | **0.356** | **78.1%** | −0.075 | 3.7% |
| All | **0.349** | **78.3%** | −0.099 | −0.2% |

The outcome is worse still than under the binary rules, even though this specification exits the market far less often (between sixteen and thirty-two months, against fifty-three for the regime strategy). The full portfolio ends the period with a cumulative return of −0.2%, that is, no gain whatsoever over eight years.

### 5.8.2 Diagnosis: the backward-looking nature of the signal

That a strategy exiting the market on few occasions should nonetheless destroy almost all returns suggests those exits are concentrated in the wrong months. The data confirm this: in the full portfolio, months spent out of the market averaged +3.07% returns, against +0.09% for months spent invested. Of the eight best months in the evaluation period, the strategy missed seven, including April 2020 (+14.4%), November 2020 (+12.4%) and October 2022 (+9.7%).

The underlying reason is that regime probability is a backward-looking rather than forward-looking signal. Table 14 reports correlations between the filtered bear-regime probability and market returns at various leads and lags.

**Table 14:** Correlation between the filtered bear-regime probability at $t$ and market returns at different points in time. Full portfolio, evaluation period.

| | **Correlation** |
|---|---|
| Probability($t$) vs. return($t-3$) | $-0.217$ |
| Probability($t$) vs. return($t-2$) | $-0.217$ |
| Probability($t$) vs. return($t-1$) | $-0.200$ |
| Probability($t$) vs. return($t$) | $+0.098$ |
| Probability($t$) vs. return($t+1$) | $+0.089$ |

The signal correlates negatively with *past* returns, since bear-regime probability rises after declines, but its correlation with next month's return is essentially nil and positive in sign, the opposite of what anticipating the market would require. Regime probability therefore summarises what has already happened, with no appreciable predictive content.

This is not an implementation flaw. A regime classification built mainly on lagged market returns must observe several anomalous realisations before concluding that the state has changed; by the time it accumulates sufficient evidence, much of the episode has passed. We should be careful not to generalise this to regime-switching models as a class: specifications incorporating leading indicators or time-varying transition probabilities may well possess predictive content that our backward-looking classification lacks. The limitation we document applies to the specification used here. Since the sharpest rebounds in financial markets follow immediately after collapses, the signal tends to recommend exiting precisely before the recovery. April 2020 is the clearest illustration: the model identifies the bear regime after the March collapse and the strategy leaves the market just before the best month of the period.

As a benchmark for what is attainable, the *smoothed* probability, which incorporates information from the whole sample and is therefore unusable in real time, identifies months averaging +0.25% returns against +1.12% for the filtered version, the two agreeing on classification in only 69% of cases. Even with access to future information the model could not cleanly isolate unfavourable periods.

### 5.8.3 Recursive asset selection

A further objection is that these strategies are applied to portfolios aggregating all products, when one might instead restrict investment to those ETFs displaying the most favourable behaviour. If some products did possess timing ability, concentrating the portfolio in them could improve results.

Testing this requires an essential precaution. Selecting products on coefficients estimated over the full sample would introduce obvious look-ahead bias: at the time of investing, which products would prove most favourable was unknown. We therefore implement recursive selection: in each month $t$ we estimate coefficients for all eighty ETFs using information up to $t-1$ only, select the ten with the best coefficient under the criterion considered, and invest in them on an equally weighted basis over the following month, repeating the procedure each period. We consider two selection criteria, the timing coefficient and alpha, and net out the costs of the resulting turnover, which runs at 7–8% monthly.

**Table 15:** Recursive selection of the ten ETFs with the best estimated coefficient, with monthly reselection using past information only. Annualised Sharpe ratios and cumulative returns net of costs.

| **Strategy** | **Sharpe** | **Cumulative** |
|---|---|---|
| Equally weighted portfolio of all 80 (buy and hold) | 0.349 | 78.3% |
| Ten best by $\gamma$ (buy and hold) | 0.264 | 58.9% |
| Ten best by $\alpha$ (buy and hold) | **0.433** | **100.5%** |
| Ten best by $\gamma$ + regime rule | −0.075 | 10.0% |
| Ten best by $\alpha$ + regime rule | 0.027 | 20.4% |

Three readings follow. First, selecting on the timing coefficient worsens results relative to the full portfolio (58.9% against 78.3%). This is consistent with section 5.6: if timing coefficients are mostly indistinguishable from chance, selecting the highest ones amounts to selecting the products favoured by sampling noise, behaviour that does not persist.

Second, selecting on alpha does outperform the equally weighted portfolio, both in cumulative return (100.5% against 78.3%) and Sharpe ratio (0.433 against 0.349). We present this as an exploratory secondary finding, clearly separate from the main timing question, and it deserves caution: it is a single specification, the difference between the ratios has not been formally tested, and in a predominantly rising market, selecting on recent alpha may capture the persistence of the growth style rather than selection ability as such. It is also a *security selection* strategy rather than a timing one, and therefore outside the scope of this study.

Third, and most relevant for our research question, applying the timing rule damages both selected portfolios, just as it damaged the full portfolio. The alpha-based strategy sees its cumulative return fall from 100.5% to 20.4% once the regime rule is imposed. The problem lies not in the composition of the portfolio on which timing is applied, but in timing itself.

### 5.9 Summary of results

Our evidence can be summarised in four points. First, the smart beta ETFs analysed deliver negative abnormal returns, robust to specification, period and multiple testing correction, and particularly pronounced among value and dividend styles. Second, the evidence of market and volatility timing, seemingly solid under conventional inference, does not survive multiple testing correction. Third, that evidence is also unstable across volatility specifications and subperiods. Finally, and consistently, no timing strategy proves profitable out of sample.

## 6 Discussion and conclusions

### 6.1 Interpreting the negative abnormal returns

Our most robust finding is the pervasive negative alpha. Its interpretation differs substantially from what would apply in the analysis of actively managed funds.

In active management, a negative alpha is usually read as an absence of selection skill: the manager fails to identify undervalued securities to a degree sufficient to cover the fees charged. The products we analyse make no claim to discretionary selection. Their negative alpha is consistent with the cost of accessing factor exposure: management fees, index replication costs, rebalancing frictions and valuation differences. We should not attribute it exclusively to those costs, however. Benchmark misspecification, omitted factors, imperfect correspondence between an ETF's methodology and the factor model selected, and dynamic factor exposures may all contribute. The safer statement is that the estimates are consistent with implementation costs and fees, while model and benchmark limitations may also play a part.

The ordering across families reinforces this reading. Value and dividend portfolios, whose strategies are directly captured by the HML and RMW factors of the five-factor model, display the most negative alphas and do so almost universally among their constituents: every ETF in both families shows a significant negative alpha under conventional inference. This is precisely what one would expect if these products add nothing beyond factor exposure and their differential return reduces to cost. The growth family, which departs further from the linear factor combination considered, shows less negative and less systematic alphas.

This finding should not be read as a criticism of the products. An ETF that faithfully replicates a risk premium at low cost serves its purpose; a negative alpha does not signal product failure but rather that the factor model adequately captures what the product offers. The practical implication is that investors should evaluate these vehicles on their ability to deliver efficient exposure, not on their ability to generate abnormal returns.

## 6.2 The asymmetry between alpha and timing results

Before turning to the individual findings, one contrast deserves to be placed at the centre of the discussion, since it organises everything that follows. Alpha and timing coefficients behave in fundamentally different ways when subjected to scrutiny.

The negative alpha survives every test we apply. It holds across all three factor specifications, across both subperiods, and, crucially, under multiple testing correction, where twenty of the eighty products remain significant. Timing coefficients survive none of them. They reverse sign between subperiods, they depend on how volatility is measured, they cannot be cleanly separated from volatility exposure, and they all but vanish under multiplicity correction.

This asymmetry is itself informative. Abnormal underperformance appears to be a persistent and statistically robust characteristic of these products, whereas the apparent timing patterns behave as one would expect of estimates driven by sampling variation rather than by a structural feature of the funds. The fact that both sets of coefficients come from the same regressions, on the same data, under the same inference procedure, makes the contrast difficult to attribute to methodological artefact.

## 6.3 Apparent timing and the multiple testing problem

Our second central result is methodological. Under conventional inference, the analysis identified seemingly solid evidence of unfavourable timing: between fifteen and twenty-seven products with significant coefficients depending on the specification. Multiple testing correction through a block bootstrap reduces that count to zero or one.

The magnitude of this discrepancy warrants reflection. The critical threshold moves from the conventional $-1.96$ to values near $-3.3$. Put differently, when eighty assets are examined simultaneously, the probability that one of them displays a $t$-statistic below $-3$ by pure chance is substantial. Reporting the number of assets clearing the conventional threshold, standard practice in applied work, leads in this setting to badly mistaken conclusions.

This connects directly to Kosowski *et al.* (2006) and Fama and French (2010) in the mutual fund literature. The lesson is analogous: much of the apparent evidence of skill, or in our case of systematic timing behaviour, cannot be distinguished from sampling variation once tested against the distribution expected under the null. This is not the same as demonstrating that the underlying effects are zero, and we are careful not to make that stronger claim.

The internal coherence of our evidence is worth noting. The instability found in the robustness checks (coefficients switching sign across subperiods, results depending on how volatility is measured) is not a problem with the data or the specification but the expected manifestation of an absent underlying phenomenon. If systematic timing does not exist, the coefficients measuring it have no reason to behave stably.

## 6.4 The role of volatility

One intermediate result deserves specific comment. Adding the volatility timing term absorbs a substantial part of the apparent directional timing: under the five-factor model with momentum, the ten growth products with significant timing coefficients fall to none once the volatility term enters.

The explanation is natural. The Treynor-Mazuy quadratic term takes its largest values precisely when market returns are extreme in absolute terms, that is, during high-volatility periods. Both regressors therefore capture related phenomena, with a sample correlation of 0.568.

Section 5.5.2 allows us to be more precise about the mechanism. The loss of significance stems from two simultaneous factors: the timing coefficient contracts substantially, indicating a genuine reassignment of the effect towards the volatility term, and the standard error rises appreciably, reflecting the difficulty of separating two correlated regressors. We stress that this correlation does not compromise model identification, since variance inflation factors remain moderate and comparable to those of routinely used factors, but it does reduce the precision with which the two effects can be distinguished.

The implication is twofold. Classic timing tests, applied in isolation, may attribute to directional timing what is in fact sensitivity to market volatility. And more fundamentally, the data do not permit a sharp discrimination between the two phenomena, since both manifest themselves during the same periods of market turbulence.

## 6.5 Differences across styles and market states

The family-level analysis shows that Growth is the only family whose timing behaviour differs systematically across market states, with a positive coefficient in the bull regime and a negative one in the bear regime. This pattern is consistent with the construction of growth indices, oriented towards high-beta securities whose exposure amplifies during advances.

As noted in section 5.4.2, this does not amount to timing ability: a genuine market timer would display favourable coefficients in both states, whereas what we observe is procyclicality. Moreover, this evidence was obtained under conventional inference and has not been subjected to multiple testing correction, and the aggregate timing evidence vanishes under that correction. Prudence therefore suggests reading the pattern as a suggestive descriptive regularity rather than as established evidence.

## 6.6 No economic exploitation

The out-of-sample evaluation confirms this reading from a different angle. None of the strategies built on regime or volatility signals outperforms a passive position, in any portfolio. The deterioration is considerable: in the growth portfolio, the regime-based strategy destroys roughly 89% of the cumulative return attainable through a passive position.

The mechanism is transparent. The signals recommend leaving the market in more than half the months evaluated. In a period characterised by a high average market premium, the cost of standing aside far outweighs the benefit of avoiding unfavourable phases, particularly when the signals do not identify them with sufficient precision. Transaction costs, though moderate, compound an already unfavourable outcome.

Section 5.7.2 allows us to be precise that this failure does not reflect a poor choice of decision rule. A more elaborate specification, weighting expected returns in each state by their probability, performs even worse despite exiting the market far less often. The reason is that regime probability is an essentially backward-looking signal: it correlates with past returns but not with future ones. This reflects a limitation inherent to state-detection models, which require accumulating evidence before identifying a regime change, so that the signal arrives once the episode is already well advanced. Since the largest rebounds follow immediately after collapses, the strategy tends to leave the market on the eve of recovery.

Nor does restricting the portfolio to the best-performing products alter the conclusion. A recursive selection of the ten ETFs with the highest timing coefficients, carried out at each date using past information only, underperforms the full portfolio, confirming that these coefficients lack persistence. And while an analogous selection based on alpha does outperform the equally weighted portfolio (a matter of security selection rather than timing, and one that may reflect the persistence of the growth style over the sample), imposing the timing rule damages that portfolio just as much. The obstacle lies not in which assets are selected but in the decision to enter and exit the market.

We stress that this exercise was designed to respect the information available at each point in time: parameters estimated in a rolling window, lagged signals, and filtered rather than smoothed probabilities. A less careful design, using smoothed probabilities or contemporaneous regressors, would have produced artificially favourable results.

## 6.7 Limitations

Several limitations deserve explicit mention.

The first is the sample period. The 2013–2026 window features an average market premium close to 13% annually and no prolonged bear markets comparable to 2000–2002 or 2007–2009. Although it includes episodes of considerable stress (the 2020 pandemic and the 2022 bear market), its predominantly bullish character limits how far our conclusions generalise, particularly regarding the evaluation of timing strategies, which are systematically penalised in such environments.

The second concerns sample composition. Requiring a complete return history excludes products created or liquidated during the period, which avoids survivorship bias but restricts the analysis to funds already established in 2013. In addition, the small size of some families forced us to group them, precluding specific analysis of potentially interesting styles such as low volatility or momentum.

The third relates to the tests employed. Return-based timing tests are subject, as Jiang *et al.* (2007) note, to limitations that holdings-based analysis would overcome. Access to portfolio composition data would have allowed us to test directly whether index rules generate exposure variation.

Finally, the design of the strategies evaluated is deliberately simple: a binary rule for entering and leaving the market. More elaborate strategies, modulating exposure continuously or combining both signals, might yield different results, though the absence of statistical evidence of timing after multiple testing correction makes it unlikely that such an extension would change our substantive conclusions.

## 6.8 A synthesis of the evidence

Table 16 draws the main findings together, reporting for each what conventional inference suggests, what survives multiple testing correction, and what the out-of-sample evaluation shows.

**Table 16:** Synthesis of the main findings across levels of evidence. Monthly frequency, eighty ETFs.

| **Finding** | **Conventional inference** | **After multiplicity correction** | **Out-of-sample evidence** |
|---|---|---|---|
| Abnormal returns ($\alpha$) | Negative in all portfolios; 65 of 80 ETFs significant; no positive case | Holds: 20 of 80 remain significant, concentrated in Value and Dividend | Not directly tested; alpha-based selection outperforms (exploratory) |
| Market timing ($\gamma$) | Negative throughout; 15–20 ETFs significant depending on specification | Does not hold: 0–1 remain significant | Regime strategies underperform buy-and-hold in every portfolio |
| Volatility exposure ($\delta$) | Negative throughout; 21–27 ETFs significant | Does not hold: 0–1 remain significant | Volatility strategies underperform buy-and-hold in every portfolio |
| State dependence | Growth differs significantly across regimes (15 of 22 ETFs) | Not submitted to correction; treated as descriptive | Regime signal shows no predictive content for next-month returns |
| Temporal stability | — | Alpha stable across subperiods; $\gamma$ reverses sign | Consistent with absence of an exploitable pattern |

Read across the rows, the table makes the central asymmetry immediate: the alpha result strengthens as the evidentiary standard rises, while the timing results weaken and eventually disappear.

## 6.9 Conclusions

The central message of this thesis is not simply that smart beta ETFs display no market timing. Stated that way, the conclusion would be both less accurate and less interesting than what the evidence supports. What we document is that conventional return-based timing tests generate a substantial body of apparently persuasive evidence, and that this evidence proves unstable across specifications and subperiods, cannot be cleanly separated from volatility exposure, largely disappears once multiple testing is accounted for, and has no demonstrated economic value in real time.

Our answer to the research question is therefore negative, but what matters is less the result than how it is reached. The analysis shows that the conclusion depends critically on the level of statistical stringency applied. Under conventional inference, the standard practice of counting how many assets display significant coefficients, this study would have concluded that unfavourable timing is widespread. Once multiple testing is accounted for, that evidence disappears entirely. The same body of data admits two opposing readings depending on the rigour of the test employed, and only the second withstands scrutiny.

Against this fragility, one result holds across every specification, period and level of stringency: the negative abnormal return, which quantifies the cost of accessing factor exposure and is particularly pronounced among value and dividend styles.

A practical implication follows for investors. The smart beta ETFs analysed should be judged on their ability to deliver efficient and inexpensive exposure to particular risk premia, not on any capacity to adapt that exposure to market conditions. Attempts to build timing strategies upon them, even using sophisticated econometric models and statistically grounded signals, proved systematically counterproductive over the period evaluated.

Methodologically, our work illustrates two considerations of general scope. The first is how severely conventional inference misleads when numerous hypotheses are tested simultaneously: a critical threshold moving from $-1.96$ to around $-3.3$ transforms the conclusions entirely. The second is that economic evaluation must respect the information genuinely available at the moment of decision; using smoothed probabilities or contemporaneous regressors would have produced artificially favourable results without this being apparent from their presentation.

Three natural extensions suggest themselves. Extending the sample period backwards to include prolonged bear markets would test whether our conclusions hold in environments less favourable to a passive stance. Using portfolio holdings data would allow a direct test of whether index rules generate exposure variation, overcoming the limitations of return-based tests. And extending the analysis to smart beta products in other markets would assess how far our results generalise.

# References

# References

[1] Andersen, T.G. and Bollerslev, T. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review*, 39(4), 885–905.

[2] Ang, A. and Bekaert, G. (2002). International asset allocation with regime shifts. *Review of Financial Studies*, 15(4), 1137–1187.

[3] Becker, C., Ferson, W., Myers, D.H. and Schill, M.J. (1999). Conditional market timing with benchmark investors. *Journal of Financial Economics*, 52(1), 119–148.

[4] Bollen, N.P.B. and Busse, J.A. (2001). On the timing ability of mutual fund managers. *Journal of Finance*, 56(3), 1075–1094.

[5] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3), 307–327.

[6] Busse, J.A. (1999). Volatility timing in mutual funds: Evidence from daily returns. *Review of Financial Studies*, 12(5), 1009–1041.

[7] Carhart, M.M. (1997). On persistence in mutual fund performance. *Journal of Finance*, 52(1), 57–82.

[8] Fama, E.F. and French, K.R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1), 3–56.

[9] Fama, E.F. and French, K.R. (2010). Luck versus skill in the cross-section of mutual fund returns. *Journal of Finance*, 65(5), 1915–1947.

[10] Fama, E.F. and French, K.R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, 116(1), 1–22.

[11] Ferson, W.E. and Schadt, R.W. (1996). Measuring fund strategy and performance in changing economic conditions. *Journal of Finance*, 51(2), 425–461.

[12] Goetzmann, W.N., Ingersoll, J. e Ivković, Z. (2000). Monthly measurement of daily timers. *Journal of Financial and Quantitative Analysis*, 35(3), 257–290.

[13] Grinblatt, M. and Titman, S. (1994). A study of monthly mutual fund returns and performance evaluation techniques. *Journal of Financial and Quantitative Analysis*, 29(3), 419–444.

[14] Guidolin, M. and Timmermann, A. (2008). Size and value anomalies under regime shifts. *Journal of Financial Econometrics*, 6(1), 1–48.

[15] Hamilton, J.D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2), 357–384.

[16] Henriksson, R.D. (1984). Market timing and mutual fund performance: An empirical investigation. *Journal of Business*, 57(1), 73–96.

[17] Henriksson, R.D. and Merton, R.C. (1981). On market timing and investment performance II: Statistical procedures for evaluating forecasting skills. *Journal of Business*, 54(4), 513–533.

[18] Jegadeesh, N. and Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *Journal of Finance*, 48(1), 65–91.

[19] Jensen, M.C. (1968). The performance of mutual funds in the period 1945–1964. *Journal of Finance*, 23(2), 389–416.

[20] Jiang, W. (2003). A nonparametric test of market timing. *Journal of Empirical Finance*, 10(4), 399–425.

[21] Jiang, G.J., Yao, T. and Yu, T. (2007). Do mutual funds time the market? Evidence from portfolio holdings. *Journal of Financial Economics*, 86(3), 724–758.

[22] Kosowski, R. (2006). Do mutual funds perform when it matters most to investors? Mutual fund performance and risk in recessions and expansions. Working paper, INSEAD.

[23] Kosowski, R., Timmermann, A., Wermers, R. and White, H. (2006). Can mutual fund "stars" really pick stocks? New evidence from a bootstrap analysis. *Journal of Finance*, 61(6), 2551–2595.

[24] Künsch, H.R. (1989). The jackknife and the bootstrap for general stationary observations. *Annals of Statistics*, 17(3), 1217–1241.

[25] Lo, A.W. and MacKinlay, A.C. (1990). Data-snooping biases in tests of financial asset pricing models. *Review of Financial Studies*, 3(3), 431–467.

[26] Newey, W.K. and West, K.D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3), 703–708.

[27] Perez-Quiros, G. and Timmermann, A. (2000). Firm size and cyclical variations in stock returns. *Journal of Finance*, 55(3), 1229–1262.

[28] Sharpe, W.F. (1966). Mutual fund performance. *Journal of Business*, 39(1), 119–138.

[29] Sullivan, R., Timmermann, A. and White, H. (1999). Data-snooping, technical trading rule performance, and the bootstrap. *Journal of Finance*, 54(5), 1647–1691.

[30] Treynor, J.L. and Mazuy, K. (1966). Can mutual funds outguess the market? *Harvard Business Review*, 44(4), 131–136.

[31] Turtle, H.J. and Zhang, C. (2012). Time-varying performance of international mutual funds. *Journal of Empirical Finance*, 19(3), 334–348.

# A The block bootstrap procedure

This appendix formalises the multiple testing correction applied in Section 5.6.

## A.1 Notation and the testing problem

Let $i = 1, \dots, N$ index the ETFs, with $N = 80$, and $t = 1, \dots, T$ the time periods, with $T = 158$ at monthly frequency. For each asset consider the general specification

$$r_{i,t} = \alpha_i + \sum_{k=1}^{K} \beta_{i,k}\, f_{k,t} + \gamma_i\, \Psi_t + \delta_i\, VOL_t + \varepsilon_{i,t}, \tag{15}$$

where $r_{i,t}$ is the excess return, $f_{k,t}$ the $K$ risk factors, $\Psi_t$ the timing term and $VOL_t$ the volatility term.

Let $\theta_i \in \{\alpha_i, \gamma_i, \delta_i\}$ denote the parameter of interest and $t_i = \hat{\theta}_i / \hat{\sigma}(\hat{\theta}_i)$ its HAC $t$-statistic. Conventional inference rejects $H_{0,i} : \theta_i = 0$ when $|t_i| > 1.96$, which controls the individual error rate at $\Pr(|t_i| > 1.96 \mid H_{0,i}) = 0.05$.

The relevant hypothesis here is the joint one,

$$H_0 : \;\; \theta_i = 0 \quad \text{for all } i = 1, \dots, N, \tag{16}$$

and the object to be controlled is the family-wise error rate,

$$\text{FWER} = \Pr\Big( \min_{i \le N} t_i < c \;\Big|\; H_0 \Big), \tag{17}$$

where $c$ is the critical value. Under $H_0$ and for $N = 80$ correlated statistics, $\Pr(\min_i t_i < -1.96 \mid H_0) \gg 0.05$, so the conventional threshold does not control (17). The distribution of $\min_i t_i$ is unknown in closed form, since the $t_i$ are cross-sectionally dependent through common factor exposure, and is therefore approximated by bootstrap.

## A.2 Resampling scheme

Let $\ell = 6$ denote the block length in months and $m = \lceil T/\ell \rceil$ the number of blocks required. For replication $b = 1, \dots, B$ with $B = 1{,}000$, draw

$$s_j^b \overset{iid}{\sim} \mathcal{U}\{1, \dots, T - \ell + 1\}, \qquad j = 1, \dots, m, \tag{18}$$

and define the resampled index sequence

$$\mathcal{I}^b = \Big[ s_1^b, \dots, s_1^b + \ell - 1, \; \dots, \; s_m^b, \dots, s_m^b + \ell - 1 \Big]_{1:T}, \tag{19}$$

truncated to length $T$. Writing $\mathcal{I}^b = (\tau_1^b, \dots, \tau_T^b)$, the moving-block scheme of Künsch (1989) preserves the dependence structure within blocks, and hence the autocorrelation and volatility clustering of $\{\varepsilon_{i,t}\}$, which an *iid* permutation would destroy.

## A.3 Algorithm

Step 1. Estimation under the null. For each $i$, estimate the restricted model containing only the risk factors,

$$r_{i,t} = \sum_{k=1}^{K} \beta_{i,k}\, f_{k,t} + u_{i,t}, \tag{20}$$

imposing $\alpha_i = 0$ and excluding $\Psi_t$ and $VOL_t$. Retain the fitted values $\hat{r}^0_{i,t} = \sum_k \hat{\beta}_{i,k} f_{k,t}$ and the residuals $\hat{u}_{i,t} = r_{i,t} - \hat{r}^0_{i,t}$.

Step 2. Pseudo-samples. For each replication $b$ and each $i$, construct

$$r^b_{i,t} = \hat{r}^0_{i,t} + \hat{u}_{i,\tau^b_t}, \qquad t = 1, \ldots, T. \tag{21}$$

By construction $r^b_{i,t}$ satisfies $H_0$ in (16): it retains the factor exposure and the second-moment structure of the original series but contains neither abnormal return nor timing. The same index sequence $\mathcal{I}^b$ is applied to all $i$, which preserves the cross-sectional dependence of the residuals.

Step 3. Re-estimation. Fit the unrestricted model (15) to each $r^b_{i,t}$ by OLS with HAC standard errors and lag length $L = \lfloor 4(T/100)^{2/9} \rfloor$, obtaining

$$t^b_i = \frac{\hat{\theta}^b_i}{\hat{\sigma}(\hat{\theta}^b_i)}, \qquad i = 1, \ldots, N. \tag{22}$$

Step 4. Extremum statistic. Record

$$M^b = \min_{i \le N} t^b_i, \qquad b = 1, \ldots, B. \tag{23}$$

Step 5. Critical value and decision rule. The empirical distribution $\{M^b\}_{b=1}^{B}$ approximates that of $\min_i t_i$ under $H_0$. Let $c_{0.05}$ denote its fifth percentile,

$$c_{0.05} = \inf\Big\{c : \; B^{-1} \textstyle\sum_{b=1}^{B} \mathbf{1}\{M^b \le c\} \; \ge 0.05\Big\}. \tag{24}$$

Asset $i$ is declared significant if and only if

$$t_i < c_{0.05}. \tag{25}$$

The symmetric construction using $\max_i t^b_i$ and its 95th percentile applies to positive coefficients.

## A.4 Scope of the correction

The procedure is applied separately for each parameter $\theta \in \{\alpha, \gamma, \delta\}$ within each model specification. It therefore controls the family-wise error rate across the $N = 80$ ETFs for a given model and coefficient, and not jointly across the twenty-three specifications and all coefficients, which would define a larger family and require a more stringent threshold.

### A.5 Empirical critical values

Applying (24) yields $c_{0.05} = -3.30$ for $\alpha$ under FF5+MOM and $c_{0.05} = -3.89$ for $\gamma$ under TM-FF5+MOM, against the conventional $-1.96$. Under $H_0$, therefore, obtaining $\min_i t_i \approx -3$ with $N = 80$ correlated statistics is an ordinary outcome rather than evidence against the null.

Applying (25), twenty of the eighty $\alpha$ estimates remain significant, concentrated in the Value and Dividend families, whereas the fifteen to twenty $\gamma$ estimates and the twenty-one to twenty-seven $\delta$ estimates that clear the conventional threshold reduce to zero or one. Section 5.6 reports the full comparison.

Failure to reject under (25) does not establish $\theta_i = 0$; it indicates that the data do not distinguish $\hat{\theta}_i$ from sampling variation at the family-wise error rate implied by (17).

The procedure follows Kosowski *et al.* (2006), subsequently applied by Fama and French (2010) to the cross-section of mutual fund alphas.

# B Extended cross-sectional distribution

Tables 7 to 9 in the main text report the cross-sectional distribution of coefficients across all twenty-three monthly specifications. This appendix extends that information for the main specifications, presenting the three parameters of interest jointly in the manner of Fung, Xu and Yau (2002), so that the simultaneous behaviour of alpha, the timing coefficient and the volatility timing coefficient can be appreciated within each model.

**Table 17:** Joint cross-sectional distribution of coefficients in the main specifications. Eighty ETFs, monthly frequency. Alpha is expressed in percent per month. $N^+$ and $N^-$ report the number of positive and negative coefficients; in parentheses, how many are significant at the 5% level.

| Parameter | Mean | Q1 | Median | Q3 | $N^+(N^{+*})$ | $N^-(N^{-*})$ |
|---|---|---|---|---|---|---|
| *Panel A: FF5+MOM* | | | | | | |
| $\alpha$ | −0.359 | −0.434 | −0.370 | −0.309 | 0 (0) | 80 (65) |
| *Panel B: TM-FF5+MOM* | | | | | | |
| $\alpha$ | −0.273 | −0.353 | −0.294 | −0.205 | 3 (0) | 77 (42) |
| $\gamma$ | −0.449 | −0.720 | −0.433 | −0.183 | 6 (0) | 74 (15) |
| *Panel C: TM-FF5+MOM+VOL* | | | | | | |
| $\alpha$ | −0.346 | −0.431 | −0.387 | −0.264 | 1 (0) | 79 (48) |
| $\gamma$ | 0.070 | −0.180 | 0.152 | 0.309 | 48 (0) | 32 (0) |
| $\delta$ | −0.102 | −0.124 | −0.098 | −0.071 | 0 (0) | 80 (21) |
| *Panel D: CAPM+VOL* | | | | | | |
| $\alpha$ | −0.362 | −0.500 | −0.346 | −0.245 | 1 (0) | 79 (38) |
| $\delta$ | −0.154 | −0.191 | −0.142 | −0.078 | 0 (0) | 80 (27) |

Comparing panels B and C is particularly illustrative of the phenomenon documented in sections 5.3 and 5.5.2. Without the volatility term (Panel B), the timing coefficient displays a distribution clearly shifted towards negative values, with a median of −0.433 and seventy-four negative estimates out of eighty. Adding that term (Panel C) shifts the distribution towards positive values, with estimates splitting roughly evenly between signs and none reaching significance. At the same time, the volatility timing coefficient in that panel displays a uniformly negative distribution.

# C Sample composition

Table 18 lists all eighty ETFs analysed, grouped according to the classification used in this study. The "Original label" column reproduces the classification as provided by the data source, allowing the four reclassified products (IYY, FDN, FPX and KIE) to be identified, for the reasons set out in section 3.3.

**Table 18:** Sample ETFs by family.

| Ticker | Name | Original label |
|---|---|---|
| **Value** | | |
| IJJ | iShares S&P Mid-Cap 400 Value ETF | Value |
| IJS | iShares S&P Small-Cap 600 Value ETF | Value |
| ILCV | iShares Morningstar Value ETF | Value |
| IMCV | iShares Morningstar Mid-Cap Value ETF | Value |
| ISCV | iShares Morningstar Small-Cap Value ETF | Value |
| IUSV | iShares Core S&P U.S. Value ETF | Value |
| IVE | iShares S&P 500 Value ETF | Value |
| IWD | iShares Russell 1000 Value ETF | Value |
| IWN | iShares Russell 2000 Value ETF | Value |
| IWS | iShares Russell Mid-Cap Value ETF | Value |
| IWX | iShares Russell Top 200 Value ETF | Value |
| MGV | Vanguard Mega Cap Value ETF | Value |
| PWV | Invesco Dynamic Large Cap Value ETF | Value |
| RFV | Invesco S&P MidCap 400 Pure Value ETF | Value |
| RPV | Invesco S&P 500 Pure Value ETF | Value |
| RZV | Invesco S&P SmallCap 600 Pure Value ETF | Value |
| SCHV | Schwab U.S. Large-Cap Value ETF | Value |
| SPYV | SPDR Portfolio S&P 500 Value ETF | Value |
| VBR | Vanguard Small-Cap Value ETF | Value |
| VIOV | Vanguard S&P Small-Cap 600 Value ETF | Value |
| VOE | Vanguard Mid-Cap Value ETF | Value |
| VOOV | Vanguard S&P 500 Value ETF | Value |
| VTV | Vanguard Value ETF | Value |
| **Growth** | | |
| IJK | iShares S&P Mid-Cap 400 Growth ETF | Growth |
| IJT | iShares S&P Small-Cap 600 Growth ETF | Growth |
| ILCG | iShares Morningstar Growth ETF | Growth |
| IMCG | iShares Morningstar Mid-Cap Growth ETF | Growth |
| ISCG | iShares Morningstar Small-Cap Growth ETF | Growth |
| IUSG | iShares Core S&P U.S. Growth ETF | Growth |
| IVW | iShares S&P 500 Growth ETF | Growth |
| IWF | iShares Russell 1000 Growth ETF | Growth |
| IWO | iShares Russell 2000 Growth ETF | Growth |

*(continued)*

| Ticker | Name | Original label |
|---|---|---|
| IWP | iShares Russell Mid-Cap Growth ETF | Growth |
| IWY | iShares Russell Top 200 Growth ETF | Growth |
| MDYG | SPDR S&P 400 Mid Cap Growth ETF | Growth |
| RFG | Invesco S&P MidCap 400 Pure Growth ETF | Growth |
| RPG | Invesco S&P 500 Pure Growth ETF | Growth |
| RZG | Invesco S&P SmallCap 600 Pure Growth ETF | Growth |
| SCHG | Schwab U.S. Large-Cap Growth ETF | Growth |
| SPYG | SPDR Portfolio S&P 500 Growth ETF | Growth |
| VBK | Vanguard Small-Cap Growth ETF | Growth |
| VONG | Vanguard Russell 1000 Growth ETF | Growth |
| VOOG | Vanguard S&P 500 Growth ETF | Growth |
| VOT | Vanguard Mid-Cap Growth ETF | Growth |
| VUG | Vanguard Growth ETF | Growth |
| **Dividend** | | |
| DES | WisdomTree U.S. SmallCap Dividend Fund | Dividend |
| DHS | WisdomTree U.S. High Dividend Fund | Dividend |
| DLN | WisdomTree U.S. LargeCap Dividend Fund | Dividend |
| DON | WisdomTree U.S. MidCap Dividend Fund | Dividend |
| DTD | WisdomTree U.S. Total Dividend Fund | Dividend |
| DTN | WisdomTree U.S. Dividend ex-Financials Fund | Dividend |
| DVY | iShares Select Dividend ETF | Dividend |
| FVD | First Trust Value Line Dividend Index Fund | Dividend |
| HDV | iShares Core High Dividend ETF | Dividend |
| PFM | Invesco Dividend Achievers ETF | Dividend |
| SCHD | Schwab U.S. Dividend Equity ETF | Dividend |
| SDY | SPDR S&P Dividend ETF | Dividend |
| VIG | Vanguard Dividend Appreciation ETF | Dividend |
| VYM | Vanguard High Dividend Yield ETF | Dividend |
| **Other factors** | | |
| DEF | Invesco Defensive Equity ETF | Low Volatility |
| EES | WisdomTree U.S. SmallCap Fund | Fundamental |
| EZM | WisdomTree U.S. MidCap Fund | Fundamental |
| FTCS | First Trust Capital Strength ETF | Quality / Low Vol |
| MOAT | VanEck Morningstar Wide Moat ETF | Quality |

*(continued)*

| **Ticker** | **Name** | **Original label** |
|---|---|---|
| PKW | Invesco BuyBack Achievers ETF | Quality |
| PRF | Invesco RAFI US 1000 ETF | Fundamental |
| PRFZ | Invesco RAFI US 1500 Small-Mid ETF | Fundamental |
| QDEF | FlexShares Quality Dividend Defensive Fund | Quality |
| SPHQ | Invesco S&P 500 Quality ETF | Quality |
| SPLV | Invesco S&P 500 Low Volatility ETF | Low Volatility |
| TILT | FlexShares Morningstar US Market Factor Tilt | Multi-factor |
| USMV | iShares MSCI USA Min Vol Factor ETF | Low Volatility |
| **Other** | | |
| DSI | iShares MSCI USA ESG Focus ETF | ESG |
| PDP | Invesco DWA Momentum ETF | Momentum |
| RSP | Invesco S&P 500 Equal Weight ETF | Equal Weight |
| SUSA | iShares MSCI USA ESG Select ETF | ESG |
| **Sector / Broad market** | | |
| FDN | First Trust Dow Jones Internet Index Fund | Growth |
| FPX | First Trust US Equity Opportunities ETF | Momentum |
| IYY | iShares Dow Jones US ETF | Growth |
| KIE | State Street SPDR S&P Insurance ETF | EW |