Original Article

Linear and nonlinear predictability in investment style factors: multivariate evidence

Received (in revised form): 10th February 2017

Francesco Chincoli

is an analyst at Morgan Stanley, where he works in the structured credit trading team in London. He is specialized in hedging bespoke risks and providing investment solutions to institutional clients. His research concerns smart beta and more generally the predictability of portfolio returns.

Massimo Guidolin

is a chair professor of Finance and Econometrics at Bocconi University. Before, he was a professor at Manchester Business School and he also served as an assistant vice-president and senior policy consultant (Financial Markets) within the US Federal Reserve system. His research focuses on predictability and nonlinear dynamics in financial returns, with applications to portfolio management. He has published papers in the American Economic Review, the Journal of Financial Economics, the Review of Financial Studies, the Journal of Financial and Quantitative Analysis, and the Journal of Econometrics, among others. He serves on the editorial board of journals such as the International Journal of Forecasting, the Journal of Banking and Finance, and the Journal of Economic Dynamics and Control.

Correspondence: Massimo Guidolin, Finance Department, Bocconi University, via Roetgen 1, 20136 Milan, Italy
E-mail: massimo.guidolin@unibocconi.it

ABSTRACT This paper studies the predictive performance of multivariate models at forecasting the (excess) returns of portfolios mimicking the Market, Size, Value, Momentum, and Low Volatility factors isolated in asset pricing research. We evaluate the accuracy of the point forecasts of a number of linear and regime-switching models in recursive, out-of-sample forecasting experiments. We assess the accuracy of the models using several measures of unbiasedness and predictive accuracy, and using Diebold and Mariano's approach to test whether differences in expected losses from all possible pairs of forecast models are statistically significant. We fail to find evidence that complex statistical models are uniformly more accurate than a naïve constant expected return model for factor-mimicking portfolio (excess) returns. However, we show that it is possible to build simple portfolio strategies that profit from the higher out-of-sample predictive accuracy of forecasting models with Markov switching in conditional mean coefficients. These results appear to be independent of the forecasting horizon and robust to changes in the loss function that captures the investors' objectives.

Journal of Asset Management (2017) **18,** 476–509. doi:10.1057/s41260-017-0048-5;
published online 6 April 2017

Keywords: factor-mimicking portfolios; forecasting; Markov regime-switching models; equal predictive accuracy tests
JEL Codes: G11; G12; C32

INTRODUCTION

One of the most intriguing questions in asset pricing concerns to the key statistical and economic drivers of stock returns. Sharpe's (1964) and Lintner's (1965) CAPM argues that any differences in the expected excess returns of US stocks should be explained by different exposures to market risk, as measured by the related betas. Fama and French (1993), instead, proposed a three-factor model based on the excess return of the market portfolio, the return of a portfolio long in small stocks and short in big stocks (SMB), and the return of a portfolio long in high book-to-market stocks and short in low book-to-market stocks (HML). In fact, they provide evidence that the inclusion of market capitalization and book-to-market as additional risk factors effectively leads to a better explanation of the cross-section of average stock returns. In a subsequent paper (Fama and French, 1995), they also showed that these two indices, often referred to as Size and Value factors, are related to expected excess returns and are consistent with Merton's (1973) ICAPM. Moreover, Carhart (1997) augmented Fama–French's three-factor model by adding the Momentum factor, for which the factor-mimicking proxy is the return of a portfolio long in winner stocks and short in loser stocks, with performance sorting performed over the previous year (MOM).

SMB, HML, and MOM are, therefore, investable style portfolios created with the purpose of mimicking the Size, Value, and Momentum risk factors in asset pricing research (see Ferson *et al*, 2006). Interestingly, these have historically exhibited attractive risk–reward profiles and imperfect correlations with the overall market. For these reasons, factor-mimicking portfolios have captured the interest of asset managers and investors who seek to diversify their equity portfolios beyond the traditional value-weighted approach. Wealth allocation among style portfolios is however challenging, as it requires the estimation of the conditional moments of the corresponding returns. The success of any active investment strategy is inevitably linked to the accuracy of these estimates, which this paper seeks to evaluate and improve.

The objective of this study is to identify the statistical models that, if any, are the most accurate at forecasting the (excess) returns of the Market, Size, Value, Momentum, and Low Volatility APT-style factors. We perform a horse-race among a number of linear and regime-switching multivariate models and evaluate the accuracy of their out-of-sample point forecasts. We also assess the absolute and risk-adjusted performance of two simple portfolio strategies that rely on these forecasts to allocate wealth among the factor-mimicking portfolios. When appropriate, we consider three different forecasting horizons and the dividend yield, the term spread, and the default spread as prediction variables. We report four main empirical results.

First, we could not find evidence that the multivariate models considered are more accurate than a simple constant expected return model at forecasting excess returns on the Market, Size, Value, Momentum, and Low Volatility portfolios. The introduction of regime-switching parameterizations leads to overall superior measures of predictive accuracy, but the improvement is not large enough to be found statistically significant. Neither the prediction variables considered nor any autoregressive terms contribute to an increase in predictive performance.

Second, the Markov switching forecasting models, which display the best measures of predictive accuracy, can be used to generate profitable portfolio strategies. We implement two simple switching strategies that outperform standard allocations simply based on the sample mean of the factor-mimicking portfolios. The outperformance of the strategies is robust to changes in the investor's holding period and of the transaction costs imputed on her performance.

Third, the forecasts of the returns on the Low Volatility factor have required us to model Markov switching in autoregressive coefficients and heteroskedasticity dynamics in order to produce unbiased predictions. This outcome is likely to be due to the highly non-normal distribution of the returns of the associated factor-mimicking portfolio, which can be captured only by the most flexible models considered. This model also shows the highest predictive accuracy, but its forecasts fail to be significantly more accurate to plain vanilla sample mean forecasts.

Fourth, differences in predictive accuracy among models appear to be independent of the forecasting horizon. In particular, although the predictive power of the statistical models we have entertained generally worsens as the horizon lengthens, their relative performance is consistent across the three horizons considered. Hence, the horizon should not impact the choice of the model to use in empirical applications.

To best emphasize the nature of our approach, we provide further details on three key aspects of our methodology. The first aspect regards the forecast models that we compare. We analyze four classes of models, of which two are linear and two are regime-switching. In the former case, we estimate a constant expected return model and two vector autoregressive models, while in the latter case we specify two threshold vector autoregressive models and three Markov switching vector autoregressive models. More specifically, the two linear vector autoregressive models are different in the methodology of computation of the forecasts: one applies a direct method, while the other is based on iterated forecasts. The same is true for the two threshold models. The three Markov switching models, instead, vary in the number of regime-dependent parameters. A high number of time-varying parameters ought to increase the flexibility of the framework and allow for a superior in-sample fit. However, a better in-sample fit does not necessarily lead to superior out-of-sample predictive accuracy, which is instead the focus of our paper. Therefore, to produce a more robust assessment, along a Markov switching vector autoregressive model with full regime-switching parameters, we also consider simpler specifications in which only selected parameters, such as the mean and the variance–covariance matrix, are allowed to shift across regimes. Moreover, when selecting the most appropriate specification for each set of parameters, we use information criteria that balance the trade-off between in-sample fit and out-of-sample predictive accuracy.

The second aspect relates to the evaluation of the forecasts. We test whether forecasts are unbiased through a simple Wald test on the mean of the forecast errors and Mincer and Zarnowitz's regressions. We also compute several measures of predictive accuracy and use Diebold-Mariano tests to assess whether differences in performance between models are statistically significant. We rely on two loss functions, i.e., square and absolute, in order to offer a more robust analysis.

The third aspect involves the rationale behind the portfolio exercise. Similarly to Pesaran and Timmermann (1995), we adopt two simple strategies that rely on the *pseudo* out-of-sample estimates of excess returns to recursively adjust over time the allocation to style portfolios. Under the first strategy, we aim at maximizing the absolute excess return of a sample portfolio, while under the second strategy, our goal is to optimize the risk-adjusted return of an equally weighted allocation among style portfolios that only invests in portfolios with non-negative forecasted excess returns.[1] We consider three levels of transaction costs and different holding periods to assess the practical robustness and feasibility of our results.

This paper contributes to the existing literature in terms of the style risk factors studied and the econometric models considered. Our study is the first attempt to *jointly* forecast excess returns not only on the Market, Size, Value, and Momentum factors, but also on the Low Volatility factor. Haugen and Baker (1991, 2012) observed that low variance portfolios exhibit higher realized excess returns than high variance portfolio. The Authors compared the return of a portfolio long in low-volatility stocks and short in high-volatility stocks over the past year (LOWVOL) with the excess return of the capitalization-weighted market portfolio. The LOWVOL portfolio, which mimics the Low Volatility factor, has delivered average excess returns that are higher than the market portfolio on selected horizons. This evidence has spurred the interest of academics and practitioners. However, to our knowledge, no previous work has attempted to forecast the excess returns of the Low Volatility portfolio, neither in a univariate nor in a multivariate framework. Furthermore, we evaluate the predictive accuracy of a number of econometric models that is unprecedented in the literature on factor return predictability. Several studies prove that, for the Size, Value, and Momentum factors, models with regime shifts outperform single-state benchmarks not only in terms of in-sample fit, but also for their out-of-sample predictive accuracy (Black and McMillan, 2004; Guidolin and Timmermann, 2008a, b; Gulen *et al*, 2011; Perez-Quiros and Timmermann, 2000). Nevertheless, no previous work compares alternative models within the regime-switching class when forecasting factor-mimicking portfolio excess returns. Our analysis includes forecasts from threshold and Markov switching models of increasing complexity, thus allowing the switching variable to be either observable, as in the former case, or latent, as in the latter case. This approach, therefore, extends previous analyses and lays the foundations for more comprehensive comparisons. In fact, we make the strong assumption that the market excess return is the variable driving regimes in threshold models, which may explain why these models rank rather poorly when compared to Markov switching ones.[2]

Our research question is of primary importance to both researchers and practitioners in modern asset pricing and portfolio selection. A multi-index model allows the user to make factor bets. If one believes that a factor will change in a direction above that anticipated by the market, then she may want to place a bet by increasing your exposure to that factor. This can be done holding a portfolio with a sensitivity to a given factor is larger than the market index, for instance. As discussed by Ang (2014), this is important when the goal is to turn passive factor investing into an active strategy.[3] Our findings imply that such strategies may be subject to stronger limitations than commonly believed, because in spite of our best efforts we are only able to detect slim pocket of factor-mimicking predictability.

The rest of the paper is organized as follows. Second section reviews the literature on factor return predictability to which this study contributes. Third section describes the multivariate econometric models that are used in our comparisons. Fourth section presents the data and the key results of our analysis, as well as potential extensions and limitations. Final section concludes.

LITERATURE REVIEW

The academic research on stock return predictability has followed two distinct paths. On the one hand, a branch of research aims at detecting prediction variables that are able to predict stock returns (Rapach *et al*, 2005), including style returns. In this regard, some financial and macroeconomic variables, such as the dividend yield and the term spread, appear to be useful in the forecasts of excess market returns, especially in-sample (Campbell and Cochrane, 2000; Campbell and Yogo, 2005; Cochrane, 2008; Rapach and Wohar, 2006). On the other hand, regardless of the set of explanatory variables considered, predictability may require nonlinear features to be modeled (Guidolin and Timmermann, 2006). In particular, there is no doubt that regime-switching models lead to a better fit of the in-sample dynamics among asset returns. However, evidence of their out-of-sample forecasting performance is mixed.

Research on factor return predictability follows a similar pattern: it focuses either on prediction variables or on the statistical modeling of nonlinear predictive patterns. Ilmanen (2011) strongly endorses the former set of efforts. Collecting and expanding previous research on the Size, Value, and Momentum factors, he provides rational and behavioral explanations for the dynamics of the returns of the factor-mimicking portfolios. In particular, he emphasizes the existence of firm-specific and macroeconomic drivers that are likely to affect the performance of each style portfolio. Consistently, the literature mainly concentrates on univariate forecast models that incorporate such factor-specific drivers as prediction variables. For instance, Zakamulin (2013, 2014) argues that the Size premium can be predicted by a linear regression based on lagged returns, default spread, and term premium. Considering the Value factor, Cohen *et al* (2003) find forecasting power in the value spread, which they define as the difference in book-to-market ratio between a typical value stock and a typical growth stock. Chen *et al* (2008), instead, estimate the expected value premium from long-run growth rates and expected dividend-price ratios. Chordia and Shivakumar (2002) conclude that Momentum profits are explained by cross-sectional differences in conditional expected returns, which can be in turn predicted by macroeconomic variables.[4] Perez-Quiros and Timmermann (2000) were, instead, the first researchers to apply Markov switching models to style factors. They demonstrate the outperformance of heteroskedastic Markov switching models in producing point forecasts of the returns of the Size factor. Similarly to Perez-Quiros and Timmermann (2000), Gulen *et al* (2011) use a two-state, heteroskedastic regime-switching model with time-varying transition probabilities and several macroeconomic predictors in order to forecast excess returns of the Value factor.[5] Moreover, Cooper *et al* (2004), Kim *et al* (2014), and Wang and Xu (2015) show that regime-switching models are capable of producing accurate forecasts for the Momentum factor.

Several papers have engaged in comparisons of predictions of different multi-index risk factors. The literature is, however, scarce when dealing with multivariate forecast models. Lynch (2000) adopts a vector autoregressive model in order to implement optimal asset allocation on Value- and Size-sorted portfolios. He shows that their excess returns are predictable, and this predictability leads to a better risk-return ex-ante performance. Stivers and Sun (2010) forecast returns on the Value and Momentum factors using return volatility as a key predictor. Guidolin and Timmermann (2008a, b), instead, prove that regime-switching models dominate single-state benchmarks in out-of-sample forecasting experiments on the Size and Value factors. Angelidis and Tessaromatis (2014) and Sarwar *et al* (2015) have recently extended their study by introducing the Momentum factor in the analysis. However, no previous work seems to have considered multivariate models that include the Low Volatility factor.

Our paper is therefore the first to evaluate the predictive accuracy of econometric models at forecasting excess returns of the Market, Size, Value, Momentum, and Low Volatility factor portfolios. Our results, therefore, extend previous studies in APT-style factor forecastability based on the comparison between linear and regime-switching models, and should be interpreted as conditional on the multivariate framework of the analysis.

EMPIRICAL METHODOLOGY

We compare the performance of four classes of multivariate models at forecasting excess returns of factor-mimicking portfolios. We consider a constant expected return model, vector autoregressive models, threshold vector autoregressive models, and Markov switching models. Furthermore, we evaluate both direct and iterated forecasts in the vector autoregressive and the threshold vector autoregressive models.

The first two models are standard linear models. The constant expected return (CER) model may be considered the most naïve among all econometric models. It is based on the assumption that asset returns are independently and identically distributed with constant mean and variance, and, consequently, that there is no predictability beyond the sample mean of the data. In fact, one forecasts simply using the sample mean. The vector autoregressive (VAR) model introduces linear predictability from lagged returns and prediction variables. We assume that neither endogenous nor exogenous variables generate contemporaneous effects, and we estimate the parameters of each equation by Generalized Least Squares (GLS).[6] Furthermore, we estimate iterated and direct forecasts from the VAR. Iterated forecasts are based on the estimation of a one-period ahead model and the iteration of its forecasts for the desired number of periods, with forecasts plugged to replace future values of the endogenous variables. Direct forecasts, instead, use a horizon-specific model in which the explanatory variables are lagged by a desired number of periods. While direct forecasts treat predictors as exogenous variables, iterated forecasts require assumptions on the process followed by the prediction variables. Consistently with standard practice in the literature, we assume that predictors follow the same VAR model of factor-mimicking portfolio returns. This additional assumption explains why, theoretically, direct forecasts should be less affected by model misspecification, while iterated forecasts should be more efficient when the model is correctly specified (see Marcellino *et al*, 2006).

Linear models, which are static in their parameters over the entire sample, fail to capture that the relationship among variables may be unstable and dynamic. If this is the case, the CER and VAR models are likely to be misspecified. Regime-switching models try to cope with this shortcoming: they acknowledge the presence of regimes and allow parameters to vary across them. Furthermore, regime shifts can generate many nonlinear effects by mixing the conditional distributions under the different regimes. As a result, regime-switching models are able to capture the typical features of financial time series, including fat tails, skewness, heteroskedasticity, and time-varying correlations, which are overlooked by single-state, linear models (Ang and Timmermann, 2012). The threshold vector autoregressive (TVAR) and the Markov switching vector autoregressive (MSVAR) models extend linear VARs by introducing regime-specific parameters. In their most general specification, the stochastic variable S_t drives regime shifts and, thus, the parameters that characterize the time series:

$$\boldsymbol{y}_t = \boldsymbol{\mu}_{S_t} + \sum_{j=1}^{p} \mathbf{A}_{jS_t}\boldsymbol{y}_{t-j} + \boldsymbol{u}_t, \qquad (1)$$

where $\boldsymbol{u}_t \sim \mathrm{IID}(0, \boldsymbol{\Omega}_{S_t})$, $\boldsymbol{y}_t$ is a $N \times 1$ vector of random variables, and $\left\{\mathbf{A}_{1k}, \mathbf{A}_{2k}, \ldots, \mathbf{A}_{pk}\right\}_{k=1}^{K}$ are pK, $N \times N$ matrices of autoregressive coefficients.

The key difference between the two classes of models is the rationale underlying regime switches. More specifically, in threshold models, regimes are determined by the value of an observable variable (or combinations thereof), also known as threshold variable (see Tong, 1983), while, in Markov switching models, they are generated by a latent variable with Markov structure (see Hamilton, 1990). In the following, we focus on the methodology of forecasting that we adopt for threshold and Markov switching models.

Threshold vector autoregressive model

In a threshold vector autoregressive model (TVAR) (Tong, 1983), S_t assumes K values depending on the value of a *threshold variable* z at time $t-d$:

$$S_t = \begin{cases} 1 & \text{if } z_{t-d} \leq z_1^* \\ 2 & \text{if } z_1^* < z_{t-d} \leq z_2^* \\ \vdots & \vdots \\ K & \text{if } z_{K-1}^* < z_{t-d} \end{cases}, \qquad (2)$$

where $\{z_1^*, z_2^*, \ldots, z_{K-1}^*\}$ are $K-1$ thresholds and the positive integer d is the delay with which the threshold variable operates. The threshold variable can be a single variable or a combination of many variables. Furthermore, it can be endogenous or exogenous. In the former case, the model is referred to as *Self-Exciting* Threshold Autoregressive (SETAR). While the threshold variable is observable, threshold values are unobservable. Therefore, $\{z_1^*, z_2^*, \ldots, z_{K-1}^*\}$ have to be estimated in order to infer which regime prevails at any given point in time.

The TVAR model in (1) has K sets of parameters $\{\boldsymbol{\mu}_k, \mathbf{A}_{1k}, \mathbf{A}_{2k}, \ldots, \mathbf{A}_{pk}, \boldsymbol{\Omega}_k\}$, one for each regime, and the threshold values $\{z_1^*, z_2^*, \ldots, z_{K-1}^*\}$ to be estimated. A two-step procedure is usually employed for this purpose. The first step consists of the estimation of the threshold values $\{z_1^*, z_2^*, \ldots, z_{K-1}^*\}$, which are the source of the nonlinearity in the model. Afterward, the second step involves the estimation of the parameters of the resulting K VARs through Conditional Least Squares. The key issue relates to the first step, i.e., the estimation of the threshold values. We adopt the procedure adopted by Tsay (1998), which is a multivariate extension of the results in Chan (1993) and Hansen (2000). We focus on the special case of $K=2$ (two regimes and one threshold), $p=1$ (only one lag) and no exogenous variables, but the procedure can be easily generalized. With two regimes, we can write the TVAR model as:

$$\boldsymbol{y}_t = \begin{cases} \boldsymbol{\mu}_1 + \mathbf{A}_{11}\boldsymbol{y}_{t-1} + \boldsymbol{\Omega}_1^{1/2}\boldsymbol{\upsilon}_t & \text{if } z_{t-d} < z^* \\ \boldsymbol{\mu}_2 + \mathbf{A}_{12}\boldsymbol{y}_{t-1} + \boldsymbol{\Omega}_2^{1/2}\boldsymbol{\upsilon}_t & \text{if } z_{t-d} \geq z^* \end{cases}, \qquad (3)$$

where $\boldsymbol{\upsilon}_t \sim \mathrm{IID}(0, \mathbf{I}_N)$ is a white noise and $\boldsymbol{\Omega}_{S_t}^{1/2}$ can be obtained through a Cholesky decomposition of the matrix $\boldsymbol{\Omega}_{S_t}$. We assume the threshold variable z_t to be covariance stationary and continuous and estimate the parameters of the model $\left\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \mathbf{A}_{11}, \mathbf{A}_{12}, \boldsymbol{\Omega}_1^{1/2}, \boldsymbol{\Omega}_2^{1/2}, z^*, d\right\}$ in the following two-step procedure. In the first step, for given z^* and d, the model reduces to two separate VARs, depending on the values of the threshold variable. The two VARs can then be estimated individually by GLS. Among the other estimates, we obtain two sum of squares residuals (SSR), one for each equation: $\mathrm{SSR}_1(z^*, d)$ and $\mathrm{SSR}_1(z^*, d)$.[7] We define the total sum of squares of the residuals of the model in Eq. (3) as:

$$\mathrm{SSR}(z^*, d) = \mathrm{SSR}_1(z^*, d) + \mathrm{SSR}_2(z^*, d).$$

In the second step, the conditional least squares of z^* and d are:

$$\left(\hat{z}^*, \hat{d}\right) = \underset{z^*, d}{\operatorname{argmin}}\, \mathrm{SSR}(z^*, d).$$

Finally, the estimates of the parameters $\hat{\boldsymbol{\mu}}_k$, $\hat{\mathbf{A}}_{1k}$ and $\hat{\Omega}_k$ are a function of $\hat{z}^*$ and $\hat{d}$, and can be easily computed through Conditional Least Squares. We define the following set of K dummy variables $\{D_{1t}(\cdot), D_{2t}(\cdot), \ldots, D_{Kt}(\cdot)\}_{t=1}^{T}$:

$$D_{1t}(z_{t-d}) = \begin{cases} 1 & z_{t-d} \leq z_1^* \\ 0 & \text{otherwise} \end{cases},$$
$$D_{2t}(z_{t-d}) = \begin{cases} 1 & \text{if } z_1^* < z_{t-d} \leq z_2^* \\ 0 & \text{otherwise} \end{cases}, \ldots, D_{Kt}(z_{t-d}) = \begin{cases} 1 & \text{if } z_{K-1}^* < z_{t-d} \\ 0 & \text{otherwise} \end{cases},$$

and rewrite the model in (1) in the following way:

$$\boldsymbol{y}_t = \sum_{k=1}^{K} D_{kt}(z_{t-d}) \left(\boldsymbol{\mu}_k + \sum_{j=1}^{p} \mathbf{A}_{jk} \boldsymbol{y}_{t-j} \right) + \boldsymbol{u}_t, \tag{4}$$

where in addition to Eq. (1), $D_{kt}(z)$ is the kth dummy variable of the set above.[8] On a forecasting horizon $h \geq 1$, assuming $d \geq h$, we can obtain the following *direct forecast*:

$$\mathrm{E}_t[\boldsymbol{y}_{t+h}] = \sum_{k=1}^{K} D_{k(t+h)}(z_{t+h-d}) \left(\hat{\boldsymbol{\mu}}_k + \hat{\mathbf{A}}_{hk} \boldsymbol{y}_t + \hat{\mathbf{B}}_{hk} \boldsymbol{x}_t \right). \tag{5}$$

where $\boldsymbol{x}_t$ contains exogenous predictors and $d \geq h$, which implies that z_{t+h-d} and, thus, $D_{k(t+h)}(z_{t+h-d})$ are known.

In order to perform an *iterated forecast*, instead, we must treat all the variables as endogenous. On a forecast horizon $h \geq 1$, assuming $d = 1$ and $p = 1$, the one-step ahead forecast $\hat{\boldsymbol{y}}_{t+1}$ is:

$$\mathrm{E}_t[\boldsymbol{y}_{t+1}] = \sum_{k=1}^{K} D_{k(t+1)}(z_t) \left(\hat{\boldsymbol{\mu}}_k + \hat{\mathbf{A}}_{1k} \boldsymbol{y}_t \right).$$

However, when performing a multi-step ahead forecast, the situation is rather different, because z_{t+1} is unknown (and, consequently, is a random variable). More precisely, considering the forecast at time $t+2$:[9]

$$\mathrm{E}_t[\boldsymbol{y}_{t+2}] = \sum_{k=1}^{K} \hat{p}_{k(t+2)} \left(\hat{\boldsymbol{\mu}}_k + \hat{\mathbf{A}}_{1k} \sum_{k=1}^{K} D_{k(t+1)}(z_t) \left(\hat{\boldsymbol{\mu}}_k + \hat{\mathbf{A}}_{1k} \boldsymbol{y}_t \right) \right),$$

where $p_{k(t+2)} = Pr\left(D_{k(t+2)}(z_{t+1}) = 1 | z_t\right) = Pr\left(z_{k-1}^* < z_{t+1} \leq z_k^*\right)$. $\hat{p}_{k(t+2)}$ is, therefore, a function of z_{t+1}, which is a random variable. We can simulate the values of z_{t+1} in order to estimate $\hat{p}_{k(t+2)}$. In particular, assuming that z is the ith endogenous variable in the TVAR model, it follows that:

$$z_{t+1} = \sum_{k=1}^{K} D_{k(t+1)}(z_t) \left(\boldsymbol{\mu}_k^{(i)} + \mathbf{A}_{1k}^{(i)} z_t \right) + \varepsilon_{t+1},$$

where the superscript *(i)* indicates the ith row of the corresponding matrix. We generate a great number of random values for ε_{t+1} from the normal distribution and estimate $\mathrm{E}_t(z_{t+1} | \varepsilon_{t+1})$ for each of them. The estimated values are ordered and assigned to a regime depending on how they rank with respect to the estimated thresholds $\left\{\hat{z}_1^*, \hat{z}_2^*, \ldots, \hat{z}_{K-1}^*\right\}$. Finally, $\left\{\hat{p}_{k(t+2)}\right\}_{k=1}^{K}$ can be obtained by dividing the number of estimated z_{t+1} in each regime by the total number of simulations.[10] This procedure can be applied recursively until the forecasting horizon h has been reached.[11]

Markov switching vector autoregressive model

In Markov switching models, we assume that the switching variable S_t in Eq. (1) is latent and follows a discrete, first-order, k-state, time-homogeneous, irreducible, and ergodic Markov chain process.

Markov Switching models are estimated through the Expectation–Maximization (EM) algorithm (Dempster *et al*, 1977; Hamilton, 1990). Grounded in a frequentist framework, it consists of first estimating the unknown parameters of the model by Maximum Likelihood Estimation (MLE) (Krolzig, 1997), and then, conditional on the parameter estimates, making inferences on the state variable S_t.[12] The procedure is iterated until the point estimates of the parameters converge (see Guidolin, 2012).

More specifically, it is possible to rewrite the model in its state-space form:

$$\text{Measurement:} \quad \boldsymbol{y}_t = \mathbf{Y}_t \boldsymbol{\Psi}(\boldsymbol{\xi}_t \otimes \boldsymbol{\iota}_N) + \boldsymbol{\Sigma}^*(\boldsymbol{\xi}_t \otimes \mathbf{I}_N)\boldsymbol{\varepsilon}_t \tag{6}$$

$$\text{Transition:} \quad \boldsymbol{\xi}_{t+1} = \mathbf{P}'\boldsymbol{\xi}_t + \boldsymbol{u}_{t+1}, \tag{7}$$

where $\mathbf{Y}_t$ is a $N \times (Np+1)$ vector with structure $\begin{bmatrix} 1 & \boldsymbol{y}'_{t-1} & \cdots & \boldsymbol{y}'_{t-p} \end{bmatrix} \otimes \boldsymbol{\iota}_N$, $\boldsymbol{\Psi}$ is a $(Np+1) \times NK$ matrix containing the parameters of the VAR, and $\boldsymbol{\Sigma}^*$ is a $N \times NK$ matrix containing all the K Choleski factors $\{\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \ldots, \boldsymbol{\Sigma}_K\}$ for which $\boldsymbol{\Sigma}^*(\boldsymbol{\xi}_t \otimes \mathbf{I}_N)(\boldsymbol{\xi}_t \otimes \mathbf{I}_N)'(\boldsymbol{\Sigma}^*)' = \boldsymbol{\Omega}_{S_t}$. With regard to the innovations of the two equations, $\boldsymbol{\varepsilon}_t \sim NID(0, \mathbf{I}_N)$, while $\boldsymbol{u}_{t+1}$ is a zero-mean discrete martingale difference sequence vector, uncorrelated not only with $\boldsymbol{\varepsilon}_{t+1}$, but also with $\boldsymbol{\xi}_{t-j}$, $\boldsymbol{\varepsilon}_{t-j}$, and $\boldsymbol{y}_{t-j}$ $\forall j \geq 0$.

Given the measurement and the transition equations, the EM algorithm works in the following way. Calling $\boldsymbol{\theta}$ the vector collecting all the parameters of the model, we initialize the iteration with arbitrary $\boldsymbol{\xi}_1$ and $\tilde{\boldsymbol{\theta}}^0$.[13] Conditional on these parameters, we make inference on the filtered and smoothed probabilities, respectively $\left\{\tilde{\boldsymbol{\xi}}^1_{t|t}\right\}_{t=1}^T$ and $\left\{\tilde{\boldsymbol{\xi}}^1_{t|T}\right\}_{t=1}^T$ (expectation step). Conditional on these estimated smoothed probabilities, we use appropriate first-order conditions to maximize the likelihood function, which is the joint PDF of the whole sample $\{\boldsymbol{y}_t\}_{t=1}^T$, conditional on all the information available $(\Im_T)$ (maximization step):

$$f\left(\{\boldsymbol{y}_t\}_{t=1}^T \middle| \{\tilde{\boldsymbol{\xi}}^1_t\}_{t=1}^T, \Im_T\right) = \sum_{\{\tilde{\xi}^1_t\}_{t=1}^T} \prod_{t=1}^T f\left(\boldsymbol{y}_t | \tilde{\boldsymbol{\xi}}^1_t, \Im_{t-1}\right) Pr\left(\tilde{\boldsymbol{\xi}}^1_t | \boldsymbol{\xi}_1\right). \tag{8}$$

We obtain new estimates of the parameters $\tilde{\boldsymbol{\theta}}^1$ in the maximization step. These estimates are then used to make new inference on the filtered and smoothed probabilities. The EM procedure involves the iteration of the expectation and the maximization steps until convergence.[14] Under standard regularity conditions, such as identifiability and stability, the ML estimates are consistent and asymptotically normal (Hamilton, 1990; Leroux, 1992).

We refer to the following MSIVARH in order to estimate *iterated forecasts*:

$$\boldsymbol{y}_{t+1} = \boldsymbol{\mu}_{S_{t+1}} + \mathbf{A}_{1S_{t+1}}\boldsymbol{y}_t + \boldsymbol{\Omega}^{1/2}_{S_{t+1}}\boldsymbol{\upsilon}_{t+1}. \tag{9}$$

The one-step ahead forecast is:

$$\mathrm{E}\left[\boldsymbol{y}_{t+1}|\Im_t\right] = \sum_{k=1}^K \left(\hat{\boldsymbol{\mu}}_k + \hat{\mathbf{A}}_{1k}\boldsymbol{y}_t\right)\hat{\boldsymbol{\xi}}'_{t|t}\mathbf{P}\boldsymbol{e}_k.$$

However, a multi-step ahead is less straightforward. In particular, on a horizon $h > 1$, $\{\boldsymbol{y}_{t+h-1}, \boldsymbol{y}_{t+h-2}, \ldots, \boldsymbol{y}_{t+1}\}$ are unknown and have to be modeled. Their prediction, conditional on the information set $\Im_t$, depends, in turn, on the filtered probabilities $\left\{\hat{\boldsymbol{\xi}}_{t+h-1|t}, \hat{\boldsymbol{\xi}}_{t+h-2|t}, \ldots, \hat{\boldsymbol{\xi}}_{t+1|t}\right\}$, which are autocorrelated due to regime switching.[15] We follow the suggestion of Doan *et al*, (1984), who substitute $\{\hat{\mathrm{E}}[\boldsymbol{y}_{t+h-1}|\Im_t], \hat{\mathrm{E}}[\boldsymbol{y}_{t+h-2}|\Im_t], \ldots, \hat{\mathrm{E}}[\boldsymbol{y}_{t+1}|\Im_t]\}$ for $\{\mathrm{E}[\boldsymbol{y}_{t+h-1}|\Im_{t+h-2}], \mathrm{E}[\boldsymbol{y}_{t+h-2}|\Im_{t+h-3}], \ldots, \mathrm{E}[\boldsymbol{y}_{t+1}|\Im_t]\}$. Although this procedure incorrectly assumes that state price vectors are not autocorrelated, it is customary in applied studies (Guidolin, 2012). With this approximation, we can then use a simulation procedure to model

$\{\boldsymbol{y}_{t+h-1}, \boldsymbol{y}_{t+h-2}, \ldots, \boldsymbol{y}_{t+1}\}$. More specifically, given the estimates of the parameters $\left\{\boldsymbol{\mu}_k, \mathbf{A}_{1k}, \mathbf{B}_k, \boldsymbol{\Omega}_k^{1/2}\right\}_{k=1}^{K}$, of the matrix of transition probabilities $\mathbf{P}$, and of the state price vector $\boldsymbol{\xi}_t$, and assuming that $\boldsymbol{v}_t \sim NID(0, \mathbf{I}_N)$, we can simulate $\boldsymbol{y}_{t+1}$ in the following way:

$$\tilde{\boldsymbol{y}}_{t+1}^{(i)} = \sum_{k=1}^{K} \left(\hat{\boldsymbol{\mu}}_k + \hat{\mathbf{A}}_{1k}\boldsymbol{y}_t\right)\hat{\boldsymbol{\xi}}'_{t|t}\mathbf{P}\boldsymbol{e}_k + \hat{\Omega}_{t+1|t}^{1/2}\tilde{\boldsymbol{v}}_{t+1}^{(i)},$$

where, performing n simulations, the superscript *(i)* indicates the simulation number. In particular, the matrix $\hat{\Omega}_{t+1|t}$ does not equal to the average of the covariance matrices of the K regimes weighted by the filtered probabilities, because differences in the conditional means affect higher moments, including variance, and correlations among variables. The algorithm returns the set $\left\{\tilde{\boldsymbol{y}}_{t+1}^{(i)}\right\}_{i=1}^{n}$. We can iterate the procedure for each simulation i until the horizon h is reached:

$$\tilde{\boldsymbol{y}}_{t+h}^{(i)} = \sum_{k=1}^{K} \left(\hat{\boldsymbol{\mu}}_k + \hat{\mathbf{A}}_{1k}\tilde{\boldsymbol{y}}_{t+h-1}^{(i)}\right)\hat{\boldsymbol{\xi}}'_{t|t}\mathbf{P}^h\boldsymbol{e}_k + \hat{\Omega}_{t+h|t+h-1}^{1/2}\tilde{\boldsymbol{v}}_{t+h}^{(i)}.$$

Finally, the iterated forecast of $\boldsymbol{y}_{t+h}$ is the average of the simulated values at $t+h$:

$$\mathrm{E}\left[\boldsymbol{y}_{t+h}|\Im_t\right] = \frac{1}{n}\sum_{i=1}^{n}\tilde{\boldsymbol{y}}_{t+h}^{(i)}.$$

DATA AND RESULTS

The data

We study monthly data on the excess returns on five US equity portfolios and on a standard set of predictive variables sampled over the period 1929:01–2012:12, a total of 1008 observations. The data are collected from the Center for Research in Security Prices (CRSP) and Kenneth French's data repository.[16]

The portfolios are constructed in order to mimic five risk factors in the US market: Market, Size, Value, Momentum, and Low Volatility. The market portfolio is constituted by all the NYSE, AMEX and NASDAQ listed firms, value-weighted. Its excess returns are net of the risk-free rate, which we proxy using the 1-month T-Bill rate, consistently with Guidolin and Timmermann (2008a, b). We use the SMB, HML, MOM, and LOWVOL portfolios to mimic, respectively, the Size, Value, Momentum, and Low Volatility factors. The SMB and the HML portfolios are formed on the intersection of two size portfolios and three book-to-market portfolios (Fama and French, 1993). In particular, the SMB portfolio is long in small firms and short in big firms, while the HML portfolio is long in firms with high book-to-market equity and short in firms with low book-to-market equity. The MOM portfolio is long in firms with high returns and short in firms with low returns in the last 12 months (Fama and French, 2010). Finally, we construct the LOWVOL portfolio by sorting stocks on the basis of the realized variance of their returns in the previous 12 months. The factor-mimicking portfolio is the difference between the tenth decile (lowest realized variance) and the first decile (highest realized variance), and, consequently, is long in low volatility stocks and short in high-volatility stocks.

We also use lagged values of the *dividend yield, term spread* and *default spread* as prediction variables in order to improve the accuracy of the forecasts. These predictors are standard in the general forecasting literature (Fama and French, 1988; Fama, 1990; Guidolin and Ono, 2006), and the prevailing choice when forecasting excess returns of style portfolios, due to their high predictive power (Panopoulou and Plastira, 2014). The *dividend yield* (DY) is calculated as the logarithm of the ratio between the aggregate dividends on the value-weighted CRSP index over the last 12 months and the end-of-month price of the index. The *term spread* (TMS) is defined as the difference between the CRSP long-term bond yield, which corresponds to a 20-year Treasury bond rate, and the 1-month T-bill rate. Finally, the *default spread* (DFY) is the difference between the average Moody's Bbb and Aaa seasoned corporate bonds with comparable maturities. We convert term spread and default spread to continuously compounding yield in order to make them consistent with the excess returns of the factor-mimicking portfolios.

Table 1 provides summary statistics for the data. We report data on the full sample available (1929:01–2012:12) and on the prediction window (1980:01–2012:12), but we focus on the latter sample in the analysis. The data on factor-mimicking portfolio returns display features that are consistent with the asset pricing literature. The market, SMB, HML, and MOM portfolios achieve positive average excess returns, similarly to Carhart (1997). In annualized terms, the mean excess return of the market portfolio is 5.68%,[17] which is a typical value in the literature on the equity risk premium (Fama and French, 2002), with a volatility of 16.15%.[18] SMB has a lower average return, 1.11%, with a lower volatility, 10.65%; HML has an average return of 4.74%, with a volatility of 10.57%. MOM is the most aggressive portfolio: it displays the highest average return, 6.14%, with a volatility of 16.94%. LOWVOL, instead, performs poorly on the whole sample: its average return is deeply negative (−38.33%) and, at first glance, may raise some concern on the choice of LOWVOL as an investment asset. Although surprising, this finding is consistent with the finance literature, which ascertained the existence of the Low Volatility factor only in recent times (Blitz and van Vliet, 2007). The high volatility (35.16% annualized) also supports the argument that the returns of LOWVOL may be strongly heterogeneous over time. The Jarque–Bera test, whose null hypothesis of zero skewness and zero excess kurtosis is rejected in all instances, shows that factor-mimicking portfolio returns and predictors are characterized by substantial deviation from normality. Finally, the correlation of SMB, HML, MOM, and LOWVOL returns with the market excess returns is low (never higher than 33%) and often negative (up to −50%).

These features are widely reported in the literature and support investors' increasing interest in factor-mimicking portfolios. Based on this historical data, factor-mimicking portfolios are non-suboptimal investments to the market portfolio, in terms of average excess return and volatility. Furthermore, they look even more attractive on a multivariate perspective, as their negative correlation with the market allows for the construction of diversified portfolios with reduced volatility and comparable expected return. The return of such portfolios is, however, especially challenging to predict, due to the non-normalities that characterize their underlying assets, i.e., factor-mimicking portfolios. The outcomes of our study are, therefore, aimed at finding the multivariate model that is most likely to enhance the accuracy of the predictions on factor-mimicking portfolio returns.

Empirical results

We study the performance of multivariate econometric models at forecasting monthly excess returns of the five factor-mimicking portfolios over the sample 1980:01–2012:12. The exercise is recursive and out-of-sample, as we only use information that were available at the time in which the forecast, either direct or iterated, is performed.[19] However, we rely on the entire sample (1929:01–2012:12) when making inference on the goodness of fit of the models. For this reason, the finance literature usually terms these estimates *pseudo* out-of-sample (Inoue

Table 1: Summary statistics for factor-mimicking portfolio excess returns and prediction variables

Panel A: 1929:01–2012:12

	Mean	*Median*	*SD*	*Skewness*	*Kurtosis*	*Jarque–Bera*	*LB(4)*	*LB(4)-squares*
Summary statistics								
Factor-mimicking portfolios								
Market excess returns	5.17	11.23	19.03	−0.50	9.56	1852**	20**	172**
SMB	2.20	0.84	11.03	1.29	15.53	6869**	12*	36**
HML	5.99	3.36	11.90	1.81	15.40	7010**	40**	247**
MOM	6.31	10.04	18.61	−5.19	60.28	142,338**	28**	158**
LOWVOL	−65.79	−29.57	73.46	−10.82	172.19	1221854**	189**	5
Prediction variables								
Log-Dividend yield	−3.34	−3.32	0.46	−0.35	2.84	21**	3869**	3901**
Term spread	1.70	1.74	1.28	−0.30	3.34	20**	3169**	3191**
Default spread	1.14	0.90	0.70	2.32	10.65	3361**	3490**	2890**

	Excess mkt.	*SMB*	*HML*	*MOM*	*LOWVOL*	*Log-Div. yield*	*Term spread*	*Default spread*
Sample correlations								
Factor-mimicking portfolios								
Market excess returns	1.000							
SMB	0.328	1.000						
HML	0.155	0.148	1.000					
MOM	−0.335	−0.157	−0.363	1.000				
LOWVOL	−0.339	−0.562	−0.361	0.384	1.000			
Prediction variables								
Log-dividend yield	0.049	0.034	0.028	−0.071	−0.193	1.000		
Term spread	0.058	0.102	0.019	−0.077	−0.141	−0.089	1.000	
Default spread	−0.035	0.078	0.069	−0.139	−0.360	0.400	0.307	1.000

Panel B: 1980:01–2012:12

	Mean	*Median*	*SD*	*Skewness*	*Kurtosis*	*Jarque–Bera*	*LB(4)*	*LB(4)-squares*
Summary statistics								
Factor-mimicking portfolios								
Market excess returns	5.68	12.95	16.15	−1.01	6.41	259**	5	7
SMB	1.11	−0.72	10.65	0.32	10.08	835**	8	170**
HML	4.74	3.95	10.57	0.25	4.97	68**	34**	176**
MOM	6.14	9.03	16.94	−2.39	21.36	5940**	8	13*
LOWVOL	−38.33	−31.21	35.16	−2.72	20.62	5610**	38**	8
Prediction variables								
Log-dividend yield	−3.69	−3.71	0.43	0.14	2.05	16**	1525**	1529**
Term spread	2.21	2.37	1.47	−0.85	4.00	65**	1105**	1160**
Default spread	1.12	0.98	0.48	1.67	6.34	369**	1242**	1098**

	Excess mkt.	*SMB*	*HML*	*MOM*	*LOWVOL*	*Log-Div. yield*	*Term spread*	*Default spread*
Sample correlations								
Factor-mimicking portfolios								
Market excess returns	1.000							
SMB	0.248	1.000						
HML	−0.314	−0.115	1.000					
MOM	−0.119	0.031	−0.045	1.000				
LOWVOL	−0.503	−0.516	0.129	0.318	1.000			
Prediction variables								
Log-dividend yield	0.067	−0.010	−0.082	−0.025	−0.027	1.000		

Table 1 continued

	Excess mkt.	*SMB*	*HML*	*MOM*	*LOWVOL*	*Log-Div. yield*	*Term spread*	*Default spread*
Term spread	0.045	0.125	−0.054	−0.051	−0.151	−0.012	1.000	
Default spread	−0.018	0.082	−0.089	−0.142	−0.028	0.585	0.051	1.000

Notes: The table reports a few summary statistics for monthly factor-mimicking portfolio return series, and the macroeconomic variables employed as predictors of asset returns. The sample period is 1929:01–2012:12, for Panel A, and 1980:01–2012:12, for Panel B. All returns are expressed in percentage terms. LB(*j*) denotes the *j*th order Ljung-Box statistics.
* Denotes significance at the 5% level.
** Denotes significance at the 1% level.

and Kilian, 2005).[20] Moreover, we consider three forecasting horizons, i.e., 1-month, 3-month, and 12-month, which are standard not only in the forecasting literature, but also in research on portfolio optimization.

We compare four categories of multivariate models of increasing complexity. The simplest model is the constant expected return, which serves as a benchmark for the other models, as it rules out any predictability in factor excess returns. Vector autoregressive models introduce predictability from lagged returns and predictors. In this class of models, we rely on the VARX for direct forecasts and the VAR for iterated forecasts. The key difference in the two specifications relates to predictors, which are exogenous variables in the former case, while endogenous in the latter. Furthermore, threshold and Markov switching models extend the previous specifications by allowing for regime shifts in order to capture breaks and non-normal features of factor excess returns. In particular, threshold models count numerous possible specifications, depending on the lag structure, the threshold variable(s), and the delay parameter. This study focuses on a TVARX and a TVAR for better comparability with the VARX and the VAR models. The threshold variable considered is the Market excess return, with a delay equal to the smallest lag chosen. This choice is arbitrary, but consistent with previous findings in the finance literature, which stresses the persistence of bull and bear regimes across stock returns. With regard to Markov switching models, we consider several specification. First, we extend the naïve constant expected return model by introducing regime shifts in the mean parameter (MSI model). Furthermore, we allow volatility to be regime dependent as well, thus taking heteroskedasticity into account (MSIH model). Finally, we add lagged returns and predictors to the previous specification, and obtain a complete MSIVARH model. Therefore, we test a number of Markov switching models, rather than simply focusing on the model that fits the data best. This procedure allows for easier comparability, and is more robust to misspecifications in the model selection phase. As a result, we study the predictive accuracy of the following eight models: the constant expected return model (CER); two vector autoregressive models with direct and iterated forecasts (VARX and VAR, respectively); two threshold vector autoregressive models with direct and iterated forecasts (TVARX and VAR, respectively); three Markov switching models with switching intercept, of which one has no autoregressive terms and homoskedastic errors (MSI), one has no autoregressive terms and heteroskedastic errors (MSIH), and, finally, one with lag structure, predictors and heteroskedastic errors (MSIVARH).

The rest of the section is organized as follows. In section "Forecasting performance" (*Model Selection*), we select the optimal specification for each model through information criteria. In section "Testing for differential forecasting accuracy" (*Forecast Performance*), we test whether forecasts are unbiased through simple Wald tests and

Mincer–Zarnowitz regressions, and we compare the performance of the models using six standard measures of predictive accuracy. In section "Portfolio exercises" (*Testing for differential forecasting accuracy*) we test whether such differences in predictive accuracy are statistically significant. We adopt Diebold–Mariano tests for this purpose.

The first key task in any empirical application involves the selection of the most appropriate model for the data set. Model adequacy is inevitably linked with the purpose of the analysis, as a misspecified model is unlikely to produce unbiased and accurate forecast. Therefore, for each model considered, we study the specification which fits the data best. With this regard, we use three standard criteria to sort through the candidate models: the *Akaike Information Criterion* (AIC), the *Schwartz Criterion* (SC), and the *Hannan–Quinn Criterion* (HQ). For a general model, the relative statistics can be computed as follows:

$$\mathrm{AIC} = \frac{1}{T}\left(-2\log L\left(\hat{\boldsymbol{\theta}}\right) + 2\dim\left(\hat{\boldsymbol{\theta}}\right)\right),$$

$$\mathrm{SC} = \frac{1}{T}\left(-2\log L\left(\hat{\boldsymbol{\theta}}\right) + \dim\left(\hat{\boldsymbol{\theta}}\right)\ln T\right),$$

$$\mathrm{HQ} = \frac{1}{T}\left(-2\log L\left(\hat{\boldsymbol{\theta}}\right) + 2\dim\left(\hat{\boldsymbol{\theta}}\right)\ln(\ln T)\right),$$

where $-\log L(\cdot)$ is the minimum of the log-likelihood function of the model, $\hat{\boldsymbol{\theta}}$ is the vector containing the estimates of the parameters, and T is the sample size. Each criterion selects the model that minimizes the corresponding statistic. We rely on these measures because they trade-off in-sample fit with out-of-sample forecast accuracy, which is the key focus of this study (Guidolin and Ono, 2006). Furthermore, they penalize models with a large number of parameters, and therefore privilege parsimonious specifications. The AIC, SC, and HQ criteria are well established in the literature and widely used to compare models with a different number of regimes, as the statistics do not suffer of parameter nuisance (Roeder and Wasserman, 1997).

We analyze a large number of specifications which are fitted on the whole sample (1929:01–2012:12) and vary in the number of lags, p, and regimes, K. We consider models with up to three lags ($p \leq 3$), three regimes ($K \leq 3$),[21] and with a saturation ratio higher than 20 (calculated on the full sample).[22] We look at the values of the information criteria in order to choose the best specification for each model. The AIC, the SC, and the HQ criterion sometimes point at different specifications, for instance in the case of the VARX and the VAR. In particular, the AIC prefers a greater number of lags, while the SC and the HQ criteria are more parsimonious, as they usually favor models with a smaller number of parameters. This is due to the fact that, differently from the SC and the HQ criteria, the AIC is not asymptotically consistent and is biased toward more parametrized models.[23] For this reason, we rely on the Hannan-Quinn criterion and select the VARX(1,1) and the VAR(2). The selection of the TVARX(2,1,1) and the TVAR(2,2) follows the same logic. With regard to Markov switching models, instead, the decision is straightforward, as the three criteria do not diverge. Consequently, we use the MSI(3), the MSIH(3), and the MSI-VARH(3,1) models in the analysis.

Interestingly, the optimal number of regimes in threshold and Markov switching models seems to be different. Information criteria deem threshold models with more than two regimes to be suboptimal, but support the choice of richly parametrized Markov switching models. With this regard, they are especially useful to balance the trade-off implied in the selection of the appropriate number of regimes. In particular, models with many regimes are more likely to fit complex dynamics and non-normal features of the time series. However, more states demand more parameters. The higher number of parameters results in higher standard errors of the estimates, which are likely adversely affect the predictive accuracy of the model, especially if the saturation ratio is low. The sample available in our study, which includes 1008 monthly observations, is unfortunately too small to consider Markov switching models based on more than three regimes.

Forecasting performance

We study out-of-sample, point forecasts of the excess returns of the five style portfolios as follows. We recursively estimate the eight multivariate models on an expanding window of monthly observations, which starts from 1929:01 to 1979:12 and proceeds up to 2012:12. The last sample is, therefore, 1929:01–2012:11. Then, we use the models to make forecasts in the resulting window 1980:01–2012:12. In particular, each model, in every month t of the forecasting window, leads to three predictions, conditionally on the information available at time $t-1$, $t-3$, and $t-12$. Eventually, for the 5 factor portfolios, we obtain 3 estimates of returns for 396 months, coming from 8 different models, for a total of 47,520 predictions. For each point forecast, we define the forecast error from model $\mathcal{M}_i$, at time t, on a horizon h, and on factor j, as:[24]

$$\hat{e}^{j,\mathcal{M}_i}_{t,t+h} = r^j_{t+h} - \hat{r}^{j,\mathcal{M}_i}_{t,t+h},$$

where r^j_{t+h} is the actual return of factor j at time $t+h$, and $\hat{r}^{j,\mathcal{M}_i}_{t,t+h}$ is the forecast from model $\mathcal{M}_i$ for time $t+h$, conditional on the information available at time t.

In this Section, we introduce three analyses on the forecast errors. We first test whether forecast models are unbiased through a simple Wald test on the Forecast Error Bias. Furthermore, we rely on the Mincer and Zarnowitz (1969) regression to corroborate and extend the previous results. Finally, we evaluate six measures of predictive accuracy in order to compare the predictive accuracy of the models with respect to each portfolio and time horizon considered.

A forecast model $\mathcal{M}_i$ is unbiased on a horizon h if:[25]

$$\mathrm{E}\left(e^{j,\mathcal{M}_i}_{t,t+h}\right) = r^j_{t+h} - \mathrm{E}\left(r^{j,\mathcal{M}_i}_{t,t+h}\right) = 0, \quad \forall t.$$

We use the Forecast Error Bias (FEB), which is defined as the mean of the estimated forecast errors, as an estimator for the expected prediction error. Therefore, for any $\mathcal{M}_i$, h, and j, we test whether the FEB is statistically different from 0. If this is true, there is empirical evidence that model $\mathcal{M}_i$ is biased in forecasting returns of factor j, on a horizon h. Table 2 (Panel A) collects the outcome of the tests. The forecast models generally produce unbiased forecasts for the excess returns of the market, SMB, HML, and MOM portfolios, with few exceptions. In particular, the TVAR(2,2) model seems to be biased in forecasting excess returns on the 3-month and the 12-month horizon for the majority of factor-mimicking portfolios. However, the MSIVARH(3,1) stands out as the only model to be unbiased in the prediction of the excess returns of LOWVOL at all horizons.

We also rely on Mincer–Zarnowitz (1969, MZ) regressions to assess the validity of the previous results and to give more color to the analysis. We estimate the following:

$$r^j_{t+h} = \beta^j_{h,0} + \beta^j_{h,1}\hat{r}^{j,\mathcal{M}_i}_{t,t+h} + \varepsilon^{j,\mathcal{M}_i}_{t,t+h}.$$

The model $\mathcal{M}_i$ is unbiased in forecasting returns of factor j on a horizon h, if $\beta^j_{h,0}=0$ and $\beta^j_{h,1}=1$.[26] Therefore, we test the two assumptions individually, through Wald t tests, and jointly, through F-tests. We perform the regression for each combination of model $\mathcal{M}_i$, factor j, and horizon h, and display the results of the F-test in Table 2 (Panel B). The MZ regressions lead to a stricter selection, as the hypothesis that, simultaneously, $\beta^j_{h,0}=0$ and $\beta^j_{h,1}=1$ is strongly rejected in a greater number of cases. From the MZ regressions, the CER, the MSI(3), and the MSIH(3) are likely to be the

Table 2: The results of the tests that we used to verify whether the forecast models are unbiased

	CER	*VARX(1,1)*	*VAR(2)*	*TVARX(2,1,1)*	*TVAR(2,2)*	*MSI(3)*	*MSIH(3)*	*MSIVARH(3,1)*
Panel A: Test of Unbiasedness								
Excess market								
$h = 1$	0.8683	0.1687	0.3011	0.1724	0.8152	0.7524	0.8330	0.3049
$h = 3$	0.8678	0.2177	0.2947	0.1230	0.0006**	0.8387	0.8548	0.2676
$h = 12$	0.8657	0.0190*	0.2706	0.0973	0.0000**	0.8560	0.8547	0.2983
SMB								
$h = 1$	0.4929	0.8907	0.9560	0.6517	0.4589	0.7130	0.6806	0.9083
$h = 3$	0.4918	0.6906	0.9822	0.8370	0.2303	0.5426	0.4814	0.1658
$h = 12$	0.4873	0.5013	0.9894	0.5813	0.0000**	0.5026	0.5100	0.0938
HML								
$h = 1$	0.3635	0.4628	0.5843	0.9726	0.6393	0.6886	0.5932	0.2556
$h = 3$	0.3623	0.6010	0.5659	0.3081	0.2288	0.4303	0.3605	0.3813
$h = 12$	0.3556	0.3441	0.6467	0.2335	0.0000**	0.3752	0.3799	0.3758
MOM								
$h = 1$	0.7871	0.1411	0.1190	0.1009	0.0452*	0.5078	0.5486	0.0304*
$h = 3$	0.7873	0.2281	0.1167	0.5638	0.0158*	0.7152	0.7730	0.0008**
$h = 12$	0.7887	0.9831	0.0981	0.6938	0.1906	0.7652	0.7681	0.0004**
LOWVOL								
$h = 1$	0.0000**	0.0189*	0.0973	0.0103*	0.5143	0.0001**	0.0013**	0.6673
$h = 3$	0.0000**	0.0934	0.0882	0.0104*	0.0102*	0.0000	0.0000**	0.2001
$h = 12$	0.0000**	0.0820	0.0701	0.0829	0.0000**	0.0000**	0.0000**	0.0685
Panel B: MZ Regression								
Excess market								
$h = 1$	0.0948	0.0097**	0.0000**	0.0000**	0.0000**	0.8749	0.4002	0.0000**
$h = 3$	0.0579	0.0000**	0.0000**	0.0000**	0.0000**	0.9762	0.4201	0.0297*
$h = 12$	0.0709	0.0325*	0.0000**	0.0000**	0.0000**	0.0864	0.9501	0.4331
SMB								
$h = 1$	0.2812	0.0388*	0.0000**	0.0000**	0.0000**	0.7279	0.0000**	0.0000**
$h = 3$	0.2520	0.1012	0.0000**	0.5503	0.0000**	0.2835	0.7797	0.0000**
$h = 12$	0.3889	0.0021**	0.0000**	0.0001**	0.0000**	0.3970	0.1252	0.0965
HML								
$h = 1$	0.2556	0.0002**	0.0000**	0.0000**	0.0000**	0.6620	0.8454	0.0000**
$h = 3$	0.0441*	0.0001**	0.0000**	0.0412*	0.0000**	0.6934	0.1497	0.0000**
$h = 12$	0.0026**	0.0008**	0.0000**	0.0004**	0.0000**	0.0029**	0.2956	0.0004**
MOM								
$h = 1$	0.2582	0.0043**	0.0011**	0.0000**	0.0000**	0.6057	0.8148	0.0000**
$h = 3$	0.2192	0.0520	0.0014**	0.0000**	0.0000**	0.4625	0.9590	0.0017**
$h = 12$	0.3823	0.4511	0.0003**	0.0504	0.0187*	0.3537	0.4858	0.0002**
LOWVOL								
$h = 1$	0.0000**	0.0000**	0.0000**	0.0000**	0.0000**	0.0000**	0.0000**	0.0000**
$h = 3$	0.0000**	0.0000**	0.0000**	0.0000**	0.0000**	0.0000**	0.0000**	0.0000**
$h = 12$	0.0000**	0.0000**	0.0000**	0.0000**	0.0000**	0.0000**	0.0000**	0.0000**

Notes: Panel A displays the *p-values* of a Wald test on the mean of the forecast errors of each model, factor and horizon; we verify whether such means equal to 0. Panel B shows the *p-values* of the Mincer and Zarnowitz (1969, MZ) regression; using a joint F-test, we assess whether the expected value of the forecast errors is 0. In both cases, if the null hypothesis is rejected, then there is evidence that the forecast model is biased.
* Denotes significance at the 5% level.
** Denotes significance at the 1% level.

only models capable of producing unbiased forecasts on the excess returns of the market, SMB, HML, and MOM portfolios, with few exceptions. Moreover, the F-tests shows that there may not be an unbiased model for LOWVOL among the ones considered, as even the MSIVARH(3,1) appear to be biased on the three horizons. Surprisingly, unbiasedness does not appear to be correlated with the forecasting horizon, given that there are several cases in which the bias appear for $h = 1$ or $h = 3$, but not $h = 12$. The adjusted-R^2 of the regressions can also provide some indication on the predictive accuracy of unbiased models. We do not report results, but unbiased models look quite inaccurate: the CER, the MSI(3), and the MSIH(3) model, although unbiased for all factor portfolios except LOWVOL, have incredibly low R^2s (often smaller than 1%).

Table 3: Overview of the forecasting performance: best three predictive models according to alternative criteria

	Excess market	*SMB*	*HML*	*MOM*	*LOWVOL*
Panel A					
RMSFE					
$h=1$	1. MSI	1. MSI	1. MSIH	1. MSIH	1. MSI
	2. CER	2. VARX	2. MSI	2. MSI	2. CER
	3. MSIH	3. CER	3. CER	3. VARX	3. MSIVARH
$h=3$	1. MSI	1. TVARX	1. TVARX	1. MSI	1. MSIVARH
	2. CER	2. MSIH	2. MSIH	2. MSIH	2. MSI
	3. MSIH	3. CER	3. MSI	3. CER	3. CER
$h=12$	1. MSIH	1. MSI	1. MSIH	1. VARX	1. MSIVARH
	2. MSI	2. CER	2. MSI	2. CER	2. MSI
	3. CER	3. MSIH	3. CER	3. MSI	3. MSIH
Bias					
$h=1$	1. CER	1. TVAR	1. CER	1. TVAR	1. MSIVARH
	2. MSIH	2. CER	2. VARX	2. MSIVARH	2. TVAR
	3. TVAR	3. TVARX	3. VAR	3. TVARX	3. VAR
$h=3$	1. CER	1. MSIVARH	1. TVAR	1. MSIVARH	1. TVAR
	2. MSIH	2. MSIH	2. TVARX	2. TVAR	2. MSIVARH
	3. MSI	3. CER	3. CER	3. VAR	3. VARX
$h=12$	1. CER	1. MSIVARH	1. TVAR	1. MSIVARH	1. TVAR
	2. MSI	2. CER	2. TVARX	2. VAR	2. MSIVARH
	3. MSIH	3. MSI	3. VARX	3. TVAR	3. TVARX
Forecast variance					
$h=1$	1. MSI	1. MSI	1. MSIH	1. MSIH	1. MSI
	2. CER	2. VARX	2. MSI	2. MSI	2. CER
	3. MSIH	3. CER	3. CER	3. VARX	3. MSIVARH
$h=3$	1. MSI	1. TVARX	1. TVARX	1. MSI	1. MSI
	2. CER	2. MSIH	2. MSIH	2. VAR	2. CER
	3. MSIH	3. CER	3. MSI	3. VARX	3. MSIH
$h=12$	1. VARX	1. MSI	1. MSIH	1. VARX	1. MSI
	2. MSIH	2. CER	2. MSI	2. CER	2. CER
	3. MSIVARH	3. MSIH	3. CER	3. MSI	3. MSIH
MAFE					
$h=1$	1. CER	1. TVARX	1. MSIH	1. MSIH	1. MSI
	2. MSI	2. VARX	2. MSI	2. MSI	2. MSIVARH
	3. MSIH	3. MSI	3. CER	3. CER	3. CER
$h=3$	1. MSI	1. TVARX	1. MSIH	1. MSI	1. MSIVARH
	2. CER	2. MSIH	2. TVARX	2. CER	2. MSI
	3. MSIH	3. VARX	3. MSI	3. MSIH	3. CER
$h=12$	1. MSIH	1. MSI	1. MSIH	1. MSI	1. MSIVARH
	2. CER	2. CER	2. MSI	2. CER	2. MSI
	3. MSI	3. MSIH	3. CER	3. MSIH	3. MSIH
MPFE					
$h=1$	1. MSI	1. MSI	1. TVAR	1. TVAR	1. MSIH
	2. CER	2. MSIH	2. VAR	2. TVARX	2. MSI
	3. MSIH	3. CER	3. VARX	3. MSIVARH	3. CER
$h=3$	1. MSIH	1. TVARX	1. TVAR	1. TVAR	1. MSI
	2. CER	2. MSIH	2. VAR	2. MSIVARH	2. CER
	3. MSI	3. VARX	3. TVARX	3. VARX	3. MSIH
$h=12$	1. CER	1. TVARX	1. VAR	1. TVAR	1. MSI
	2. MSI	2. VARX	2. MSIVARH	2. MSIVARH	2. CER
	3. MSIH	3. TVAR	3. VARX	3. VAR	3. MSIH
Success ratio					
$h=1$	1. CER	1. MSIH	1. MSIH	1. CER	1. MSIVARH
	2. MSI	2. MSI	2. MSI	2. MSI	2. MSIH
	3. VARX	3. CER	3. MSIVARH	3. MSIH	3. VARX
$h=3$	1. CER	1. MSIH	1. MSIVARH	1. CER	1. TVAR
	2. MSI	2. CER	2. VARX	2. MSIH	2. MSIVARH
	3. MSIH	3. MSI	3. MSI	3. MSIH	3. VAR
$h=12$	1. CER	1. MSIH	1. MSIVARH	1. TVARX	1. TVAR
	2. MSI	2. CER	2. CER	2. MSIH	2. VARX
	3. MSIH	3. MSI	3. MSI	3. CER	3. MSIVARH

Table 3: continued

	Forecasting horizon			Factor portfolio					Total
	h = 1 (%)	*h = 3 (%)*	*h = 12 (%)*	*Excess mkt. (%)*	*SMB (%)*	*HML (%)*	*MOM (%)*	*LOWVOL (%)*	
Panel B									
CER	70	60	67	94	78	50	50	56	66
VARX(1,1)	30	20	23	11	33	28	33	17	24
VAR(2)	10	13	10	0	0	22	22	11	11
TVARX(2,1,1)	13	30	13	0	39	33	17	6	19
TVAR(2,2)	20	20	20	6	11	22	33	28	20
MSI(3)	73	60	70	89	61	67	56	67	68
MSIH(3)	57	70	63	94	67	56	56	44	63
MSIVARH(3,1)	27	27	33	6	11	22	33	72	29

Note: Panel B presents the percentage of times with which each model appears among the top three candidates in Panel A.

The latest result highlights that unbiasedness gives little indication about predictive accuracy. If a forecast model is unbiased on a horizon h, its predictions for $t+h$, conditional on time t, are, on average, centered on the actual values of the variable at $t+h$. However, there is no guarantee that unbiased forecasts are more accurate than biased ones. Therefore, in addition to the FEB, we compute five measures of predictive accuracy, which are standard in the forecasting literature (Guidolin *et al*, 2009). These measures are the Root Mean Squared of Forecast Error (RMSFE), the Forecast Error Variance (FEV), the Mean Absolute Forecast Error (MAFE), the Mean Percent Forecast Error (MPFE), and the Success Ratio (SR). In particular, for each factor-mimicking portfolio j, horizon h, and model $\mathcal{M}_i$:

$$\mathrm{RMSFE}^{j,h,\mathcal{M}_i} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\hat{e}^{j,\mathcal{M}_i}_{t,t+h}\right)^2}$$

$$\mathrm{FEV}^{j,h,\mathcal{M}_i} = \frac{1}{T}\sum_{t=1}^{T}\left(\hat{e}^{j,\mathcal{M}_i}_{t,t+h} - \bar{e}^{j,\mathcal{M}_i}\right)^2$$

$$\mathrm{MAFE}^{j,h,\mathcal{M}_i} = \frac{1}{T}\sum_{t=1}^{T}\left|\hat{e}^{j,\mathcal{M}_i}_{t,t+h}\right|$$

$$\mathrm{MPFE}^{j,h,\mathcal{M}_i} = \frac{100\%}{T}\sum_{t=1}^{T}\frac{\hat{e}^{j,\mathcal{M}_i}_{t,t+h}}{r^{j}_{t+h}}$$

$$\mathrm{SR}^{j,h,\mathcal{M}_i} = \frac{1}{T}\left(\# r^{j}_{t+h} * \hat{r}^{j,\mathcal{M}_i}_{t,t+h} \geq 0\right),$$

where T is the size of the sample, and $\# r^{j}_{t+h} * \hat{r}^{j,\mathcal{M}_i}_{t,t+h} \geq 0$ indicates the number of times in which the forecast return and the actual return have the same sign. Given the high number of statistics, we decided to collect them in Table 3 as follows: Panel A displays the top three models in the six measures for each model $\mathcal{M}_i$ and forecast horizon h, given factor j, while Panel B focuses on the percentage of times in which a model ranks among the top three candidates for each measure considered. Consistently with the MZ regressions, the CER, MSI(3), and MSIH(3) are generally the best performing models, as they appear in the top of the ranking 66, 68, and 63% of the times, respectively. This result is remarkable because it shows that, although with some exceptions, linear and nonlinear models without autoregressive terms lead to a superior performance to models with a lag structure. Consequently, it casts doubts on the effective predictive power of the dividend yield, the term spread and the default spread in a multivariate setting.[27] The good performance of the CER model is surprising and suggests that factor-mimicking portfolio excess returns may be randomly distributed around their sample mean, thus supporting the idea that they may not be predictable. This is partially consistent with general findings on stock return predictability with monthly data. Establishing a clear ranking among the CER, the MSI(3), and the MSIH(3) models is instead hard, as their measures of predictive accuracy display similar values. More specifically, the MSI(3) and the MSIH(3) models produce smaller RMSFE, MAFE, and SR than the CER model on all the horizons considered. This is mainly due to their ability to minimize the Forecast Error Variance (FEV) and is in line with the results in Guidolin *et al*, (2009), who prove the superior predictive accuracy of Markov Switching models. In addition, the MSI(3) and the MSIH(3) models perform marginally better when forecasting excess returns of SMB, HML, and MOM. Interestingly, threshold models not only are outperformed by Markov switching models in many instances, but are also unlikely to improve the forecast accuracy of their corresponding linear models, as they often rank worse in almost all measures.

The performance of the multivariate models seems to be independent of the forecast horizon. Although the accuracy of the forecasts generally worsens as the horizon lengthens, with the RMSFE and the MAFE increasing and the SR decreasing, the rankings in Table 3 are, in fact, not affected by changes in h. The rankings are, instead, very different depending on the factor being considered. For instance, the CER, the MSI(3), and the MSIH(3) model dominates when forecasting the market excess returns, for which they rank in the top class 94, 89, and 94% of the times, respectively, while their outperformance is less tangible for the HML and MOM portfolios. The Low Volatility factor-mimicking portfolio, instead, appear to be the toughest to predict: RMSFE and MAFE are 2–4 times higher for LOWVOL, while the success ratio is as low as 63%. The MSIVARH(3,1) model, which is the most flexible among the eight considered, is particularly successful in capturing its dynamics (72% of the times, against an average of 18% on the other factors). The MSIVARH(3,1) is always among the top picks for minimizing the RMSFE, the MAFE and the FEV on LOWVOL, while still producing forecasts having the smallest possible bias, as we expected from the previous findings.

In fact, the analysis has been conducted on a purely relative basis so far. We have proved the statistical unbiasedness of some forecast models and found out which ones are more accurate in recursive, out-of-sample forecasts of the excess returns of the factor-mimicking portfolios. However, the conclusion that one model is likely to perform better than another does not imply that its forecasts are good enough to be used in empirical applications. For instance, the MAFE, which gives an intuitive indication of the average forecast error, ranges between 2.0 and 3.5% for MKT-RF, SMB, HML, and MOM, while their monthly average excess return is only between 0.1 and 0.6%. The adjusted-R^2s of the MZ regressions stress the same concept, and casts further doubt on actual predictability in factor-mimicking portfolio excess returns in a multivariate setting.

Testing for differential forecasting accuracy

We have tested whether the eight forecast models are unbiased, and ranked them according to their predictive accuracy. However, a relative comparison among forecasts cannot be comprehensive: the fact that model $\mathcal{M}_i$ ranks better than model $\mathcal{M}_g$ does not imply that the former is statistically more accurate than the latter. In other words, the measures and tools used so far do not provide evidence that the difference between $\mathcal{M}_i$ and $\mathcal{M}_g$ is statistically different from 0. In this Section, we apply Diebold and Mariano's (1995, DM) test of equal predictive accuracy, which exactly serves this purpose.

The DM statistic tests whether the mean loss function values obtained from two alternative forecast models $\mathcal{M}_i$ and $\mathcal{M}_g$ are statistically different. Given a loss function

$L\left(\varepsilon_{t,t+h}^{j,\mathcal{M}_i}\right)$, we define the difference of the loss function of the two competing models in the following way:

$$\mathrm{diff}_{t,j,h}^{\mathcal{M}_i,\mathcal{M}_g} \equiv L\left(\varepsilon_{t,t+h}^{j,\mathcal{M}_i}\right) - L\left(\varepsilon_{t,t+h}^{j,\mathcal{M}_g}\right).$$

We test the null hypothesis that the mean of these differences is 0, i.e., $\mathrm{E}\left(\mathrm{diff}_{t,j,h}^{\mathcal{M}_i,\mathcal{M}_g}\right) = 0$, through the following DM statistics:

$$\mathrm{DM}_{t,j,h}^{\mathcal{M}_i,\mathcal{M}_g} = \frac{\frac{1}{T-h}\sum_{t=1}^{T} \mathrm{diff}_{t,j,h}^{\mathcal{M}_i,\mathcal{M}_g}}{\hat{\sigma}\left(\mathrm{diff}_{t,j,h}^{\mathcal{M}_i,\mathcal{M}_g}\right)},$$

where $\hat{\sigma}\left(\mathrm{diff}_{t,j,h}^{\mathcal{M}_i,\mathcal{M}_g}\right)$ is the estimate of the standard error of the loss differential between $\mathcal{M}_i$ and $\mathcal{M}_g$ and is usually computed by the Newey–West estimator.[28] The DM statistic, under the assumption that the loss differential is covariance stationary, follows an asymptotically normal distribution (Diebold and Mariano, 1995): $\mathrm{DM}_{t,j,h}^{\mathcal{M}_i,\mathcal{M}_g} \sim N(0,1)$. If the null hypothesis is rejected, then we can conclude that model $\mathcal{M}_i$ leads to an expected loss that is higher than the expected loss from model $\mathcal{M}_g$. Hence, the latter model should be preferred.

The loss function indicates the *cost* or *disutility* associated with the differences between forecast return and actual return, and is, therefore, a function of the forecast error. The appropriate loss function depends on the utility function and the decision environment of the user (Diebold and Mariano, 1995). This study is aimed at general applications; therefore, we consider two standard loss functions: square and absolute.[29] In the former case, the loss associated with each prediction is the squared error, $L\left(\varepsilon_{t,t+h}^{j,\mathcal{M}_i}\right) = \left(\varepsilon_{t,t+h}^{j,\mathcal{M}_i}\right)^2$, while in the latter case it is the absolute error, $L\left(\varepsilon_{t,t+h}^{j,\mathcal{M}_i}\right) = \left|\varepsilon_{t,t+h}^{j,\mathcal{M}_i}\right|$. The choice of these two functions is supported by Diebold and Lopez (1996), who highlight that optimal forecast errors enjoy desirable statistical properties under a squared loss function.[30] In addition, the absolute loss function, although not differentiable in $\varepsilon_{t,t+h}^{j,\mathcal{M}_i} = 0$, is more robust to outliers. The square and the absolute loss functions are symmetric, and, consequently, assign the same weight, or cost, to positive and negative errors. In empirical applications, the actual economic loss of the agent should be evaluated in order to define the appropriate function. Its utility function may, for instance, require an asymmetric loss function, thus depending on the entire distribution of forecast returns, rather than simply on their point estimate. It is also possible to specify a loss function that is different for the returns of each factor-mimicking portfolio. This may be the case of a mutual fund consisting of a pool of investors with unequal amount of wealth allotted in the strategies.

The DM test requires the loss differential be covariance stationary. Using Augmented Dickey–Fuller tests, we verify whether this assumption is empirically correct for all the combinations of prediction errors, under the square and absolute loss function. The null hypothesis of unit root is rejected at the 1% significance level in all instances. Hence, we proceed in calculating the DM statistics. Tables 4, 5, 6, 7 and 8 display the result of the DM tests for each style portfolio. The three Panels (A, B, and C) refer to the forecasting horizons considered ($h = 1, 3, 12$) and are structured in the following way. The cells *below* the main diagonal contain the p-values of the DM test under the square loss function. The null hypothesis is rejected if the MSFE of two forecast models are statistically different. Similarly, the cells *above* the main diagonal contain the p-values of the DM test under the absolute loss function. In this case, the null hypothesis is rejected if the MAFE of two forecast models are statistically different. Values lower than 5% are marked in bold. In either case, if the null hypothesis is not rejected, there is not enough statistical evidence to argue that the performance of a model is significantly better than the other.

Similarly to what we have previously noticed, the performance of the forecast models is strongly dependent on the portfolio under consideration, but it is very similar on the three horizons. For this reason, in the following, we refer to values in Panel A of the five Tables, unless stated otherwise.

Under square loss function, the CER, MSI(3), and MSIH(3) models provide significantly better performance than the other models only for the market excess returns, for which the DM statistic is significant at the 1% level. Considering the absolute loss function, this is confirmed and extended to the Momentum factor. However, the DM tests do not validate the impression that the MSI(3) and the MSIH(3) lead to superior performance to the CER model for SMB, HML, and MOM, as the three models rarely show significant differences in predictive accuracy. Furthermore, the underperformance of the TVARX(2,1,1) and, especially, of the TVAR(2,2) against Markov switching models is clear from the low p-values. Results are, however, mixed when comparing the two threshold models with their corresponding linear specification. The dominance of the MSIVARH(3,1) in forecasting excess returns of LOWVOL is remarked against all models but the CER and the MSI(3), while it outperforms the MSIH(3) only for $h = 1$. Finally, clear differences between the two loss functions can be noticed only with regard to MOM.

Portfolio exercises

The DM tests did not give clear-cut indications as to whether MSI(3) and MSIH(3) may be significantly more accurate than a simple CER model. In this Section, we demonstrate that these forecasting models can in fact be employed to support profitable real-time strategies. In the following, we present a portfolio exercise based on two simple switching strategies, which we label “Strategy 1” and “Strategy 2”. The two strategies share a common approach because, similarly to Pesaran and Timmermann (1995), they rely on *pseudo* out-of-sample forecasted returns to allocate wealth among style portfolios. Nonetheless, they differ in their investment rationale and objective.

In Strategy 1, we use the forecasting models to maximize the excess return of our sample portfolio. Therefore, at the beginning of each month, we allocate all the available wealth in the single style portfolio that, according to the selected model, is expected to yield the higher excess return in the subsequent month. In Strategy 2, we use the models to enhance the risk-adjusted return of an equally weighted allocation among style portfolios. In practice, we rely on the equally weighted portfolio, but, at the beginning of each month, we exclude any asset that is predicted to generate a negative excess return in the subsequent month according to a given model. In both strategies, we rebalance the allocation monthly, over the sample 1980:01–2012:12 and their realized excess return at time $t+1$ is:

$$R_{t+1}^{\mathcal{M}_i} = \sum_j \hat{w}_{t,t+1}^{j,\mathcal{M}_i} \cdot r_{t+1}^j = \sum_j \frac{\hat{D}_{t,t+1}^{j,\mathcal{M}_i}}{\sum_j \hat{D}_{t,t+1}^{j,\mathcal{M}_i}} r_{t+1}^j,$$

where $\mathcal{M}_i$ is the model, j is the factor portfolio index, and D is a dummy that differs according to the strategy. In Strategy 1, at time t we assign weight 1 to the portfolio that, according to model $\mathcal{M}_i$, yields the highest forecasted excess return, and weight 0 to any other portfolio:

$$\hat{D}_{t,t+1}^{j,\mathcal{M}_i} = \begin{cases} 1 & \text{if } \hat{r}_{t,t+1}^{j,\mathcal{M}_i} = \text{Max}\left\{\hat{r}_{t,t+1}^{j,\mathcal{M}_i}\right\}_j \\ 0 & \text{otherwise} \end{cases}.$$

In Strategy 2, at time t we assign an equal weight $1/n$ to each of the n factor portfolios that, according to model $\mathcal{M}_i$, generate non-negative forecasted excess return for time $t+1$, and weight 0 to the other $5 - n$ portfolios, which, in turn, generate negative predictions:

Table 4: The Diebold-Mariano equal predictive accuracy tests: Market excess returns

	CER	*VARX(1,1)*	*VAR(2)*	*TVARX(2,1,1)*	*TVAR(2,2)*	*MSI(3)*	*MSIH(3)*	*MSIVARH(3,1)*
Panel A: 1-month horizon								
CER	–	**0.010**	**0.001**	**0.004**	**0.000**	0.730	0.738	**0.029**
VARX(1,1)	0.219	–	0.060	0.150	**0.001**	**0.012**	**0.024**	0.336
VAR(2)	**0.003**	**0.007**	–	0.718	**0.006**	**0.001**	**0.001**	0.927
TVARX(2,1,1)	**0.030**	0.055	0.567	–	**0.009**	**0.005**	**0.008**	0.865
TVAR(2,2)	**0.003**	**0.006**	**0.014**	**0.013**	–	**0.000**	**0.000**	**0.029**
MSI(3)	0.793	0.200	**0.003**	**0.021**	**0.002**	–	0.925	**0.039**
MSIH(3)	0.756	0.267	**0.004**	**0.035**	**0.003**	0.716	–	**0.025**
MSIVARH(3,1)	0.072	0.138	0.858	0.594	**0.022**	0.077	0.059	–
Panel B: 3-month horizon								
CER	–	**0.000**	**0.003**	**0.001**	**0.010**	0.720	0.418	**0.025**
VARX(1,1)	**0.027**	–	0.903	0.946	0.630	**0.000**	**0.001**	**0.020**
VAR(2)	**0.005**	0.471	–	0.926	0.718	**0.003**	**0.004**	**0.028**
TVARX(2,1,1)	**0.041**	0.865	0.422	–	0.613	**0.001**	**0.002**	**0.022**
TVAR(2,2)	0.068	0.514	0.840	0.453	–	**0.009**	**0.014**	0.089
MSI(3)	0.115	**0.022**	**0.004**	**0.033**	0.061	–	0.411	**0.022**
MSIH(3)	0.683	**0.032**	**0.006**	**0.048**	0.074	0.353	–	**0.038**
MSIVARH(3,1)	0.238	0.114	**0.011**	0.139	0.134	0.189	0.262	–
Panel C: 12-month horizon								
CER	–	0.075	**0.006**	**0.001**	**0.000**	0.066	0.491	0.271
VARX(1,1)	0.634	–	**0.021**	0.116	**0.000**	0.076	0.063	**0.036**
VAR(2)	**0.005**	**0.002**	–	0.727	**0.000**	**0.006**	**0.005**	**0.003**
TVARX(2,1,1)	**0.020**	0.058	0.589	–	**0.001**	**0.001**	**0.001**	**0.007**
TVAR(2,2)	**0.003**	**0.002**	**0.005**	**0.018**	–	**0.000**	**0.000**	**0.000**
MSI(3)	0.840	0.633	**0.005**	**0.020**	**0.003**	–	0.392	0.277
MSIH(3)	0.365	0.585	**0.004**	**0.018**	**0.003**	0.361	–	0.228
MSIVARH(3,1)	0.868	0.501	**0.001**	**0.030**	**0.002**	0.867	0.763	–

Notes: The table presents *p*-values for Diebold and Mariano's (1995, DM) tests of no difference in predictive accuracy. Boldfaced *p*-values are below the 5% threshold. In each panel, in cells above the main diagonal we report *p*-values under an absolute loss function; in cells below the main diagonal we show *p*-values under a square loss function.

$$\hat{D}_{t,t+1}^{j,\mathcal{M}_i} = \begin{cases} 1 & \text{if } \hat{r}_{t,t+1}^{j,\mathcal{M}_i} \geq 0 \\ 0 & \text{otherwise} \end{cases}.$$

Transaction costs may strongly affect the two active strategies, as investors inevitably face costs, fees, and commissions any time they modify their allocation. In our exercise, we present results for three levels of transaction costs: 0, 50, and 100 basis points. We also consider their ex-ante impact on monthly allocations: we keep the weights unchanged if the increase in total excess return that we expect from a switch is insufficient to justify the cost of switching.

Table 9 presents the mean, standard deviation, and Sharpe ratio of each exercise. In Strategy 1, we recursively look for the model that can maximize portfolio excess returns. The MSIH(3) shines in the comparison, as it generates a portfolio with an average annualized excess return of 9.58%, which is more than twice the average excess return of the CER portfolio. The MSI(3) is the runner up model: its portfolio displays an average excess return that is approximately 1% lower than the MSIH(3) portfolio, but well above the average excess returns of any other model. Furthermore, the two switching models yield the portfolios with the lowest volatility and, as a result, the highest Sharpe ratio. The CER model, which selects the best performing allocation on the basis of historical mean excess returns, is suboptimal. We can derive further insight into the performance of the three models by looking at their allocation among factors (Table 10). The sample mean seems to be sufficient to recognize that HML and MOM are likely to have high excess returns. However, the CER model suggests to invest in HML more often than necessary (19% of the times). Markov switching models are more effective because

Table 5: The Diebold-Mariano equal predictive accuracy tests: SMB

	CER	*VARX(1,1)*	*VAR(2)*	*TVARX(2,1,1)*	*TVAR(2,2)*	*MSI(3)*	*MSIH(3)*	*MSIVARH(3,1)*
Panel A: 1-month forecast horizon								
CER	–	0.243	0.786	0.204	**0.000**	0.250	0.246	0.051
VARX(1,1)	0.916	–	**0.048**	0.520	**0.000**	0.483	0.083	**0.002**
VAR(2)	0.262	0.117	–	0.067	**0.000**	0.530	0.819	0.056
TVARX(2,1,1)	0.563	0.480	0.900	–	**0.000**	0.355	0.066	**0.000**
TVAR(2,2)	**0.000**	**0.000**	**0.000**	**0.001**	–	**0.000**	**0.000**	**0.000**
MSI(3)	0.520	0.941	0.209	0.533	**0.000**	–	**0.008**	**0.025**
MSIH(3)	0.116	0.105	0.994	0.869	**0.000**	0.081	–	0.077
MSIVARH(3,1)	0.052	**0.030**	0.103	**0.004**	0.202	**0.048**	0.053	–
Panel B: 3-month forecast horizon								
CER	–	0.955	0.885	0.655	**0.000**	0.919	0.156	**0.003**
VARX(1,1)	0.788	–	0.855	0.513	**0.000**	0.948	0.908	**0.027**
VAR(2)	0.405	0.443	–	0.666	**0.000**	0.889	0.811	0.144
TVARX(2,1,1)	0.545	0.399	0.385	–	**0.000**	0.648	0.755	0.067
TVAR(2,2)	**0.000**	**0.000**	**0.011**	**0.000**	–	**0.000**	**0.000**	**0.004**
MSI(3)	0.590	0.830	0.418	0.529	**0.000**	–	0.193	**0.003**
MSIH(3)	0.341	0.714	0.378	0.590	**0.000**	0.221	–	**0.001**
MSIVARH(3,1)	**0.040**	0.063	0.806	0.219	**0.011**	**0.045**	**0.033**	–
Panel C: 12-month forecast horizon								
CER	–	0.365	0.714	0.371	0.073	0.290	0.376	0.154
VARX(1,1)	**0.003**	–	0.914	0.659	0.076	0.360	0.392	0.875
VAR(2)	0.473	0.670	–	0.713	0.222	0.710	0.735	0.839
TVARX(2,1,1)	0.088	0.236	0.930	–	0.177	0.367	0.393	0.931
TVAR(2,2)	**0.030**	**0.033**	0.055	0.064	–	0.072	0.074	0.217
MSI(3)	0.436	**0.003**	0.472	0.087	**0.030**	–	0.293	0.152
MSIH(3)	0.158	**0.011**	0.485	0.097	**0.030**	0.114	–	0.175
MSIVARH(3,1)	0.324	0.364	0.572	0.259	0.052	0.323	0.398	–

Notes: The table presents *p*-values for Diebold and Mariano's (1995, DM) tests of no difference in predictive accuracy. Boldfaced *p*-values are below the 5% threshold. In each panel, in cells above the main diagonal we report *p*-values under an absolute loss function; in cells below the main diagonal we show *p*-values under a square loss function.

they recognize, *ex ante*, that MOM will be the best performing selection: MSI(3) and MSIH(3) select the MOM factor portfolio 98 and 92% of the time, respectively. Moreover, the two models effectively mitigate periods of downturn in MOM by switching allocation to HML, a more defensive portfolio. The other models, instead, generate portfolios with lower excess returns and higher volatility.

In Strategy 2 we aim at identifying style portfolios predicted to give positive excess returns to then perform a static, equally weighted allocation (EW) among them. We rely on the Sharpe ratio to measure risk-adjusted performance and compare the resulting portfolios. We can appreciate how, consistently with the literature, EW improves the Sharpe ratio of the individual factor-mimicking portfolios (to 0.6914) through a drastic reduction in volatility (6.39%). As a result, EW should be the investors' preferred method in the absence of excess return predictions. This result no longer holds once predictability is taken into account. Table 9 shows that MSI(3) generates a higher Sharpe ratio (0.7683) than EW, mainly due to a higher average excess return (4.91%). The MSIH(3) model, which was the top performer in Strategy 1, ranks below EW, as it generates a portfolio with slightly higher average excess return (4.66%) but lower Sharpe ratio (0.6147). The other models clearly underperform. A key reason is due to their excessive allocation to LOWVOL, which leads to severe drags in excess returns. MSIVARH(3,1), nevertheless, is associated with the portfolio with the highest mean excess return (6.21%). Consistently with our earlier findings, this is due to its ability to identify periods of high positive returns from LOWVOL.

Table 6: The Diebold-Mariano equal predictive accuracy tests: HML

	CER	*VARX(1,1)*	*VAR(2)*	*TVARX(2,1,1)*	*TVAR(2,2)*	*MSI(3)*	*MSIH(3)*	*MSIVARH(3,1)*
Panel A: 1-month forecast horizon								
CER	–	0.288	0.367	0.745	**0.000**	0.879	**0.042**	0.068
VARX(1,1)	0.427	–	1.000	0.445	**0.000**	0.332	**0.025**	0.257
VAR(2)	0.394	0.865	–	0.519	**0.000**	0.390	**0.049**	0.315
TVARX(2,1,1)	0.624	0.984	0.957	–	**0.000**	0.707	0.095	0.119
TVAR(2,2)	**0.002**	**0.003**	**0.003**	**0.013**	–	**0.000**	**0.000**	**0.025**
MSI(3)	0.582	0.418	0.385	0.585	**0.003**	–	**0.008**	0.080
MSIH(3)	0.083	0.127	0.123	0.291	**0.001**	0.053	–	**0.006**
MSIVARH(3,1)	0.194	0.268	0.398	0.353	**0.030**	0.220	0.080	–
Panel B: 3-month forecast horizon								
CER	–	0.342	0.417	0.599	**0.001**	0.209	**0.022**	**0.019**
VARX(1,1)	0.120	–	0.718	0.237	**0.011**	0.250	0.127	0.128
VAR(2)	0.219	0.921	–	0.284	0.169	0.362	0.237	0.406
TVARX(2,1,1)	0.583	0.206	0.166	–	**0.002**	0.724	0.919	**0.036**
TVAR(2,2)	**0.005**	0.066	0.226	**0.002**	–	**0.001**	**0.000**	0.589
MSI(3)	0.507	0.090	0.182	0.638	**0.005**	–	0.110	**0.011**
MSIH(3)	0.120	0.093	0.123	0.804	**0.002**	0.297	–	**0.008**
MSIVARH(3,1)	0.072	0.623	0.795	0.136	0.224	0.056	**0.047**	–
Panel C: 12-month forecast horizon								
CER	–	0.451	0.127	0.440	**0.000**	0.108	0.354	0.186
VARX(1,1)	0.467	–	0.565	0.782	0.074	0.439	0.420	0.528
VAR(2)	0.131	0.520	–	0.482	0.205	0.121	0.112	0.929
TVARX(2,1,1)	0.445	0.889	0.269	–	**0.028**	0.427	0.404	0.413
TVAR(2,2)	**0.001**	0.289	0.501	0.278	–	**0.000**	**0.000**	0.140
MSI(3)	0.436	0.461	0.127	0.439	**0.001**	–	0.501	0.179
MSIH(3)	0.199	0.448	0.119	0.425	**0.001**	0.321	–	0.169
MSIVARH(3,1)	0.269	0.909	0.559	0.950	0.247	0.261	0.253	–

Notes: The table presents *p*-values for Diebold and Mariano's (1995, DM) tests of no difference in predictive accuracy. Boldfaced *p*-values are below the 5% threshold. In each panel, in cells above the main diagonal we report *p*-values under an absolute loss function; in cells below the main diagonal we show *p*-values under a square loss function.

Transaction costs fail to materially affect the comparisons between the MSI(3), the MSIH(3), and the CER model under both Strategies 1 and 2. However, the performance of the other forecast models rapidly deteriorates with an increase in switching costs, as their recommended allocation relies on a more frequent rebalancing.

So far, we have calculated mean excess returns, volatility, and Sharpe ratio for the entire pseudo out-of-sample window (1980:01–2012:12). The results reported above are, therefore, conditional on this sample: they could only be replicated by an investor that followed Strategy 1 or Strategy 2 exactly between January 1980 and December 2012. We address this limitation by verifying the robustness of our analysis in different subsamples. In particular, we first compute the mean excess return and Sharpe ratio that an investor would achieve by following Strategies 1 and 2 from month t to month $t+H$, where H is the holding period of the investment. We then calculate the same statistics for an investment from month $t+H+1$ to month $t+2H$ and proceed in a similar fashion until the end of the forecasting window. We obtain a time series of non-overlapping estimates of average excess returns and Sharpe ratios for each model, given a holding period H. Finally, we compute the mean of these series and the standard error of the mean.[31] If the performance of a model portfolio is uniform over the out-of-sample window, we expect means that are similar to the full sample statistics. Furthermore, the lower the standard error of the mean, the more consistent the performance. Table 11 collects estimates for holding periods of 1, 3, and 5 years. We present the average excess returns under Strategy 1 and the Sharpe ratios from Strategy 2. We can

Table 7: The Diebold-Mariano equal predictive accuracy tests: MOM

	CER	*VARX(1,1)*	*VAR(2)*	*TVARX(2,1,1)*	*TVAR(2,2)*	*MSI(3)*	*MSIH(3)*	*MSIVARH(3,1)*
Panel A: 1-month forecast horizon								
CER	–	0.970	0.904	0.399	0.115	0.507	0.446	0.121
VARX(1,1)	0.196	–	0.680	0.125	0.117	0.485	0.407	**0.037**
VAR(2)	0.368	0.547	–	0.207	0.124	0.380	0.328	0.078
TVARX(2,1,1)	**0.022**	**0.029**	**0.038**	–	0.184	0.107	0.091	0.398
TVAR(2,2)	**0.009**	**0.036**	**0.024**	0.168	–	0.087	0.092	0.261
MSI(3)	0.792	0.164	0.331	**0.018**	**0.008**	–	0.701	**0.042**
MSIH(3)	0.652	0.141	0.285	**0.016**	**0.011**	0.720	–	**0.041**
MSIVARH(3,1)	**0.031**	0.226	0.177	0.980	0.182	**0.028**	**0.027**	–
Panel B: 3-month forecast horizon								
CER	–	0.985	0.973	0.404	0.068	0.104	0.596	0.112
VARX(1,1)	**0.020**	–	0.978	0.321	0.078	0.895	0.947	0.533
VAR(2)	0.461	0.404	–	0.370	0.083	0.894	0.940	0.573
TVARX(2,1,1)	**0.008**	0.273	0.144	–	0.192	0.371	0.378	0.664
TVAR(2,2)	**0.008**	0.179	0.100	0.501	–	0.059	0.062	0.108
MSI(3)	**0.020**	**0.011**	0.369	**0.005**	**0.005**	–	0.526	0.059
MSIH(3)	0.814	**0.019**	0.473	**0.008**	**0.009**	0.069	–	0.107
MSIVARH(3,1)	0.452	0.174	0.780	0.067	**0.016**	0.294	0.489	–
Panel C: 12-month forecast horizon								
CER	–	0.932	0.819	0.313	0.058	0.738	0.703	**0.015**
VARX(1,1)	0.178	–	0.800	0.216	0.053	0.927	0.874	0.061
VAR(2)	0.423	0.691	–	0.991	0.948	0.820	0.826	0.658
TVARX(2,1,1)	**0.039**	0.093	0.981	–	0.790	0.318	0.350	0.255
TVAR(2,2)	**0.030**	0.152	0.993	0.967	–	0.058	0.050	0.122
MSI(3)	0.409	0.175	0.422	**0.039**	**0.030**	–	0.730	**0.014**
MSIH(3)	0.353	0.209	0.438	**0.048**	**0.036**	0.301	–	**0.016**
MSIVARH(3,1)	0.345	0.905	0.757	0.603	0.491	0.341	0.366	–

Notes: The table presents *p*-values for Diebold and Mariano's (1995, DM) tests of no difference in predictive accuracy. Boldfaced *p*-values are below the 5% threshold. In each panel, in cells above the main diagonal we report *p*-values under an absolute loss function; in cells below the main diagonal we show *p*-values under a square loss function.

appreciate how, similarly to the previous results, the MSI(3) and MSIH(3) lead the rankings. Under Strategy 1, MSIH(3) posts an average excess return of 9.6% in the full sample and for an investment of 1 or 3 years. The standard error of this estimate is small: between 0.8 and 1.0%. Under Strategy 2, the Sharpe ratio from the two models actually increases from approximately 0.8 to 1.1.

Discussion

We have performed a horse-race among eight multivariate forecast models and tested the properties of their *pseudo* out-of-sample predictions of the (excess) returns of five factor-mimicking portfolios. We could not find statistical evidence that the multivariate econometrics models considered are significantly more accurate than a simple constant expected return model at forecasting excess returns on the Market, Size, Value, Momentum, and Low Volatility portfolios. However, we proved that it is possible to build simple portfolio strategies to exploit the higher out-of-sample predictive accuracy of two models with Markov switching in conditional mean coefficients. The key results of the analysis are:

1. *Among the eight forecast models considered, no one is statistically superior for all the style portfolios analyzed*. Our purpose was to find the multivariate model which is the most accurate at forecasting the excess returns of all the factor portfolios. We found, however, no clear winner among the competing model.
2. *The CER, MSI(3), and MSIH(3) models generally lead to better predictive accuracy than forecast models that include autoregressive terms*. This is especially evident in their ability to minimize the MSFE

Table 8: The Diebold-Mariano equal predictive accuracy tests: LOWVOL

	CER	*VARX(1,1)*	*VAR(2)*	*TVARX(2,1,1)*	*TVAR(2,2)*	*MSI(3)*	*MSIH(3)*	*MSIVARH(3,1)*
Panel A: 1-month forecast horizon								
CER	–	**0.000**	**0.000**	**0.000**	**0.000**	0.604	**0.006**	0.229
VARX(1,1)	**0.000**	–	**0.007**	0.672	**0.000**	**0.000**	0.717	**0.003**
VAR(2)	**0.000**	0.070	–	**0.005**	0.096	**0.000**	0.381	**0.000**
TVARX(2,1,1)	**0.000**	0.231	**0.017**	–	**0.000**	**0.000**	0.605	**0.013**
TVAR(2,2)	**0.000**	**0.001**	**0.017**	**0.000**	–	**0.000**	0.174	**0.000**
MSI(3)	**0.001**	**0.000**	**0.000**	**0.000**	**0.000**	–	**0.001**	0.163
MSIH(3)	0.129	**0.009**	**0.001**	**0.033**	**0.000**	**0.001**	–	**0.032**
MSIVARH(3,1)	0.383	**0.000**	**0.000**	**0.000**	**0.000**	0.266	**0.031**	–
Panel B: 3-month forecast horizon								
CER	–	**0.017**	**0.000**	**0.012**	**0.000**	0.201	0.890	0.753
VARX(1,1)	0.101	–	**0.000**	0.375	**0.005**	**0.013**	**0.016**	**0.002**
VAR(2)	**0.000**	**0.000**	–	**0.000**	0.151	**0.000**	**0.000**	**0.000**
TVARX(2,1,1)	**0.049**	0.446	**0.000**	–	0.057	**0.009**	**0.006**	**0.004**
TVAR(2,2)	**0.000**	**0.000**	0.576	**0.007**	–	**0.000**	**0.000**	**0.000**
MSI(3)	0.063	0.059	**0.000**	**0.024**	**0.000**	–	0.565	0.930
MSIH(3)	0.652	0.122	**0.000**	**0.045**	**0.000**	0.167	–	0.689
MSIVARH(3,1)	0.241	**0.004**	**0.000**	**0.003**	**0.000**	0.493	0.143	–
Panel C: 12-month forecast horizon								
CER	–	**0.041**	**0.019**	**0.043**	**0.006**	**0.000**	0.325	0.688
VARX(1,1)	**0.048**	–	0.115	0.546	0.233	**0.039**	**0.041**	**0.018**
VAR(2)	**0.009**	0.214	–	0.122	0.920	**0.019**	**0.019**	**0.013**
TVARX(2,1,1)	**0.041**	0.794	0.217	–	0.206	**0.042**	**0.044**	**0.020**
TVAR(2,2)	**0.004**	0.216	0.620	0.201	–	**0.006**	**0.006**	**0.004**
MSI(3)	**0.000**	**0.045**	**0.009**	**0.038**	**0.004**	–	0.394	0.765
MSIH(3)	0.277	**0.047**	**0.009**	**0.040**	**0.004**	0.194	–	0.731
MSIVARH(3,1)	0.450	**0.007**	**0.002**	**0.005**	**0.001**	0.530	0.486	–

Notes: The table presents *p*-values for Diebold and Mariano's (1995, DM) tests of no difference in predictive accuracy. Boldfaced *p*-values are below the 5% threshold. In each panel, in cells above the main diagonal we report *p*-values under an absolute loss function; in cells below the main diagonal we show *p*-values under a square loss function.

and the MAFE, but less obvious when looking at the forecast bias. The superior performance of models without a lag structure casts doubts on the effectiveness of the predictors considered, even though these represent the classical predictors in the asset pricing literature.

3. *The MSI(3) and the MSIH(3) models can be used to generate outperforming strategies both in absolute and in risk-adjusted terms.* Diebold–Mariano tests show that there is not enough statistical evidence to prove that the two Markov switching model lead to significantly more accurate forecasts. However, real-time portfolio exercise suggests a different conclusion: the two models generated portfolio allocations that clearly outperformed the corresponding benchmarks in Strategy 1 and Strategy 2.
4. *The threshold models considered fail to improve the forecasting performance of the corresponding (single-state) linear models.* This should not be interpreted as a general failure of threshold models, but rather as a shortcoming of the chosen specifications. We indeed made restrictive, a priori assumptions on the threshold variable and the delay parameter. It would be useful to test whether threshold variables different from the market excess returns may lead to more accurate predictions, similarly to Guidolin *et al*, (2009).
5. *The relative predictive accuracy of forecast models is strongly heterogeneous across factors, while generally homogeneous across the three forecast horizons.* The rankings in Table 3 show that forecast models that are top performers for some factor-mimicking portfolios are often not

Table 9: The results of the two *pseudo* out-of-sample portfolio exercises

	MKT-RF	*SMB*	*HML*	*MOM*	*LOWVOL*	*EW*
Factor-mimicking portfolios						
No transaction costs						
Mean	5.7%	1.1%	4.7%	**6.1%**	−38.3%	4.4%
SD	16.1%	10.6%	10.6%	16.9%	35.1%	**6.4%**
Sharpe ratio	0.352	0.104	0.449	0.363	−1.092	**0.691**

	CER (%)	*VARX(1,1) (%)*	*VAR(2) (%)*	*TVARX(2,1,1) (%)*	*TVAR(2,2) (%)*	*MSI(3) (%)*	*MSIH(3) (%)*	*MSIVARH(3,1) (%)*
Strategy 1								
No transaction costs								
Mean	4.5	3.1	−3.4	5.2	−5.1	8.3	**9.6**	6.7
SD	16.8	16.4	21.1	18.2	21.7	15.0	**14.5**	15.1
Sharpe ratio	0.271	0.187	−0.163	0.286	−0.236	0.557	**0.661**	0.443
Low transaction costs								
Mean	4.4	−0.5	−6.7	1.6	−9.1	8.3	**9.0**	3.6
SD	16.7	16.5	21.1	18.3	21.8	15.0	**14.6**	15.1
Sharpe ratio	0.261	−0.028	−0.319	0.088	−0.420	0.556	**0.615**	0.237
High transaction costs								
Mean	3.4	−3.0	−10.6	−1.4	−12.9	8.2	**9.0**	0.3
SD	16.2	16.6	21.1	18.5	21.8	14.9	**14.5**	15.2
Sharpe ratio	0.209	−0.181	−0.503	−0.075	−0.591	0.546	**0.623**	0.021
Strategy 2								
No transaction costs								
Mean	4.4	3.3	−0.9	3.5	−0.9	4.9	4.7	**6.2**
SD	**6.4**	9.3	12.0	11.5	10.4	**6.4**	7.6	11.0
Sharpe ratio	0.691	0.360	−0.071	0.301	−0.089	**0.768**	0.615	0.567
Low transaction costs								
Mean	4.4	1.6	−3.6	0.5	−3.2	**4.9**	4.1	4.3
SD	**6.4**	9.2	12.4	11.5	10.5	**6.4**	7.1	11.0
Sharpe ratio	0.691	0.173	−0.289	0.046	−0.302	**0.763**	0.574	0.389
High transaction costs								
Mean	4.4	0.0	−5.3	−1.4	−4.9	**4.8**	3.5	2.6
SD	6.4	9.4	12.3	11.4	10.6	**6.4**	7.0	10.9
Sharpe ratio	0.691	−0.003	−0.431	−0.121	−0.460	**0.758**	0.501	0.241

Notes All data is annualized and calculated over the sample 1980:01–2012:12. EW is a portfolio with equal weights in MKT-RF, SMB, HML, and MOM. Low transaction costs and high transaction costs indicate, respectively, 50 bps and 100 bps. Values in bold show the best result in each measure and strategy.

Table 10: The average allocation suggested by the forecasting models for each strategy and factor-mimicking portfolio

	CER (%)	VARX(1,1) (%)	VAR(2) (%)	TVARX(2,1,1) (%)	TVAR(2,2) (%)	MSI(3) (%)	MSIH(3) (%)	MSIVARH(3,1) (%)
Strategy 1								
No transaction costs								
MKT-RF	–	8.3	14.1	10.4	13.9	–	0.5	19.9
SMB	–	8.3	6.8	11.1	24.0	–	–	8.1
HML	19.2	26.8	23.2	18.9	10.9	2.3	7.6	14.9
MOM	80.8	39.9	32.3	41.4	22.2	97.7	91.9	51.8
LOWVOL	–	16.7	23.5	18.2	29.0	–	–	5.3
Low transaction costs								
MKT-RF	–	8.6	14.1	9.8	13.9	–	–	20.2
SMB	–	8.1	7.1	10.9	24.2	–	–	7.8
HML	2.8	26.8	22.7	19.2	10.6	2.5	7.8	14.1
MOM	97.2	39.9	33.1	41.9	22.0	97.5	92.2	52.5
LOWVOL	–	16.7	23.0	18.2	29.3	–	–	5.3
High transaction costs								
MKT-RF	–	8.6	14.1	10.1	13.6	–	–	20.7
SMB	–	7.6	7.3	10.6	24.2	–	–	8.1
HML	49.7	27.0	22.5	18.9	10.6	2.5	8.1	13.9
MOM	50.3	39.9	33.1	42.2	22.2	97.5	91.9	52.0
LOWVOL	–	16.9	23.0	18.2	29.3	–	–	5.3
Strategy 2								
No transaction costs								
MKT-RF	25.0	16.6	17.0	16.5	20.5	25.2	22.6	20.9
SMB	25.0	18.7	17.4	19.4	20.7	25.2	21.7	17.1
HML	25.0	27.3	24.6	24.6	20.4	25.2	29.2	22.4
MOM	25.0	27.4	27.1	28.7	25.0	24.5	26.6	34.6
LOWVOL	–	10.0	13.5	10.8	13.4	–	0.0	4.5
Low transaction costs								
MKT-RF	25.0	16.9	17.1	16.3	20.7	25.2	23.4	20.7
SMB	25.0	18.6	17.0	19.4	20.4	25.2	22.6	17.2
HML	25.0	27.2	24.5	24.8	20.4	25.2	28.3	22.2
MOM	25.0	27.3	27.6	28.8	25.0	24.5	25.7	35.2
LOWVOL	–	10.0	13.6	10.7	13.4	–	0.0	4.4
High transaction costs								
MKT-RF	25.0	17.2	17.0	16.3	20.8	25.2	23.8	21.1
SMB	25.0	17.9	16.8	19.3	20.2	25.2	23.1	17.1
HML	25.0	27.1	25.1	25.0	20.5	25.2	27.8	22.1
MOM	25.0	27.6	27.4	28.6	25.2	24.5	25.3	34.9
LOWVOL	–	10.2	13.5	10.7	13.3	–	–	4.4

Notes: Low transaction costs and high transaction costs indicate, respectively, 50 bps and 100 bps.

as good for other portfolios. The most striking example is the MSIVARH(3,1), which is almost always the top pick in the case of LOWVOL, but never the preferred choice for the other portfolios. This outcome suggests that, in empirical applications, the relative importance of the factor-mimicking portfolios is likely to be determinant in the choice of the optimal forecast model. Furthermore, we argued that the relative performance of the models is independent of the forecasting horizon because the rankings do not materially vary across the three horizons analyzed.

6. *The Low Volatility factor shows the most complex dynamics*. Table 3 suggests that the majority of models fail at adapting to the idiosyncratic features and non-normalities of the returns of LOWVOL. Predictions from the MSIVARH(3,1) clearly stand out as the most accurate according to the indicators considered. Therefore, we argue that the MSIVARH(3,1), which is the most flexible among the models considered, should be the prevailing

Table 11: The average excess return and Sharpe ratio of an investor for each combination of strategy, forecasting model and holding period

	MKT-RF	*SMB*	*HML*	*MOM*	*LOWVOL*	*EW*
Factor-mimicking portfolios						
Holding Period	Average excess return: mean					
1	5.7%	1.1%	4.7%	6.1%	−38.3%	4.4%
3	5.7%	1.1%	4.7%	6.1%	−38.3%	4.4%
5	5.9%	0.6%	4.9%	5.2%	−40.7%	4.1%
Holding period	Average Sharpe ratio: mean					
1	0.646	0.000	0.393	0.767	−1.270	1.147
3	0.487	0.132	0.433	0.585	−1.071	0.920
5	0.539	0.056	0.478	0.496	−1.203	0.867

	CER	*VARX(1,1)*	*VAR(2)*	*TVARX(2,1,1)*	*TVAR(2,2)*	*MSI(3)*	*MSIH(3)*	*MSIVARH(3,1)*
Strategy 1								
Holding period	Average excess return: mean							
1	4.5%	3.1%	−3.4%	5.2%	−5.1%	8.3%	**9.6%**	6.7%
3	4.5%	3.1%	−3.4%	5.2%	−5.1%	8.3%	**9.6%**	6.7%
5	4.4%	1.4%	−5.1%	3.5%	−5.6%	7.6%	**8.0%**	6.2%
Holding period	Average excess return: standard error of the mean							
1	0.9%	0.9%	1.1%	1.0%	1.0%	0.6%	0.8%	0.6%
3	0.7%	1.1%	1.0%	1.0%	1.0%	0.6%	1.0%	0.6%
5	1.0%	1.4%	1.5%	0.8%	1.7%	0.7%	1.0%	0.5%
Strategy 2								
Holding period	Average Sharpe ratio: mean							
1	1.147	0.682	0.194	0.619	0.075	1.189	**1.193**	0.958
3	0.920	0.504	0.114	0.518	0.093	**0.988**	0.950	0.767
5	0.867	0.342	−0.036	0.311	0.004	**0.940**	0.824	0.764
Holding period	Average Sharpe ratio: standard error of the mean							
1	0.225	0.197	0.182	0.187	0.186	0.217	0.224	0.168
3	0.257	0.263	0.229	0.237	0.217	0.220	0.275	0.147
5	0.277	0.272	0.242	0.231	0.229	0.236	0.278	0.207

Notes: Holding periods indicate an investor that entered the portfolio at any given point in time and held it for an equivalent amount of time (in years). We use non-overlapping investment periods and assume no transaction costs. Values in bold show the best result in each measure, strategy, and holding period.

choice for the Low Volatility factor. However, the DM tests do not support the significance of such results against the CER and the MSI(3) models, and Table 2 highlights contrasting results for the unbiasedness properties.

7. *Inference on the optimal model is conditional on the loss function adopted*. The square and the absolute loss functions led to negligible differences in results. However, the choice of the appropriate model to use should be conducted under a loss function that is tailored to individual preferences (i.e., the utility function) of the user. Therefore, a change in loss function may skew results in favor of specific models.
8. *In absolute terms, the forecast models are quite inefficient at forecasting factor returns.* This is evident in the high values of the RMSFE, the FEV, and the MAFE, as well as in the low R^2 of the MZ regressions.

We showed that forecasts from the eight multivariate forecast models are inaccurate, albeit mostly unbiased (point 8). We also reached the conclusion that, under square and absolute loss functions, there is no optimal forecast model for all style portfolios (point 1). This may lead to believe that a shortcoming lies in the choice of adopting a multivariate, rather than univariate, approach. Previous studies have indeed focused on forecasting the excess returns of each factor-mimicking portfolio individually, rather than jointly, see Arnott *et al*, (1989) and section "Literature review". A univariate setting is likely to lead to more accurate predictions, especially in the case of regime-switching models.[32] However, a univariate framework would not take into account the joint dynamics of factor-mimicking portfolios excess returns and consequently limit the scope of our results. The multivariate nature of this study is useful to a number of applications. For instance, any portfolio optimization involving factor-mimicking portfolios has to deal with the estimation of their expected returns, variances, and correlations.[33] In particular, Ang *et al*, (2010) show how performance evaluation and active asset allocation change when multiple style portfolios are considered. Investors and practitioners should also be interested in managing the risks embedded in their portfolios. Standard risk management models, such as the parametric Value at Risk (VaR), are based on expected returns, volatilities and correlations. The reliability of these models, consequently, depends on the accuracy of such estimates.

CONCLUSIONS

This paper shows that there may be no optimal model to predict the (excess) returns of the Market, Size, Value, Momentum, and Low Volatility factor-mimicking portfolios. We could not find enough statistical evidence to argue that sophisticated models can significantly improve over the predictive accuracy of a simple constant expected return model. However, we also found that there is value in modeling Markov switches in conditional mean coefficients when implementing simple asset allocation strategies across style portfolios. These results suggest then that forecasting (excess) returns using models that account for regimes should be a matter of key importance to investors seeking to maximize absolute or risk-adjusted returns from portfolios that time the state of asset markets.

These findings should not be taken as applicable to in all situations because they are limited to the range of models and predictors considered. Furthermore, they are conditional on the specification of a loss function, which we considered to be either square or absolute. Hence, they may not be reliable in all empirical applications where the user is adopting a tailor-cut economic loss which may be, for instance, asymmetric or different across factor portfolios. Similarly, evaluating interval forecasts rather than point forecasts may prove the dominance of one of the models that this study considered suboptimal. Finally, the multivariate framework of analysis lays the foundations for forecasting experiments that involves not only moments of higher order beyond the conditional mean (i.e., simple point forecasts), but also correlations among factors. The flexibility of regime-switching models may prove particularly useful in this respect, and it may shed light on the limitations of the constant expected return model, which has performed surprisingly well in our study.

We outline several extensions of the analysis. First, additional factors and predictors could be fruitfully considered. Although the dividend yield, the term spread, and the default spread do not contribute to the predictive accuracy of the forecast models considered, there may be other variables with more resilient explanatory power.[34] It may also be useful to evaluate the forecast accuracy of other models, such as Smooth Transition Vector Autoregressive models, or to introduce conditional heteroskedasticity in the linear specifications. Second, several aspects of the methodology could be further developed. For instance, we performed a forecasting experiment on expanding windows of data, which may favor regime-switching models against single state ones. Comparing out-of-sample forecasts on rolling windows of data may change the outcome of the analysis. In addition, we relied solely on the DM approach when evaluating differences in the losses across models. Other tests of predictive accuracy, such as the van Dijk and Franses's (2003) weighted test of equal prediction accuracy and the Giacomini and White's (2006) test, would possibly corroborate our analysis. Third, the results are strongly dependent on the measures of bias and accuracy used in the comparison. In turn, these statistics are based on a loss function, which we have assumed to be either square or absolute value. The actual cost incurred by the user could well be different. Hence, comparing results under additional loss functions should provide more information on the appropriateness of the forecast models under diverse decision making environments. Finally, we considered only one of the possible applications of our methodologies. The outcome might of course have differed in case, for instance, we have assessed the performance of the forecasting models under more sophisticated portfolio strategies. It remains an exciting strand of research to pursue additional tests of multivariate models with regimes.

ACKNOWLEDGEMENTS

We would thank one anonymous referee for helpful comments and suggestions to improve the paper.

NOTES

1. Equally weighted portfolios are widely used in practice. DeMiguel, Garlappi and Uppal (2009) show that they are often performing well in pseudo out-of-sample backtesting exercises.
2. It could be interesting to test the performance of models based on different threshold variables. The inclusion of additional models, such as smooth transition VAR models, and prediction variables, such as return dispersion (see Stivers and Sun, 2010), would be an essential step in order to check the robustness of our conclusions. We do not pursue this extension in our paper.
3. Ang (2014) discusses the so-called "Professors' Report to the Norwegian Ministry of Finance" to list four criteria for determining which factors investors should choose; a factor should be: (i) justified by academic research, (ii) have exhibited significant premiums *that are expected to persist in the future* (emphasis added), (iii) have return history available for bad times, (iv) be implementable in liquid, traded instruments. Our research impinges on point (ii) and implies that risk premia may be positive and persistent but predicting the direction of changes of such premia may be as hard as ever.
4. They also find that the value-weighted market dividend yield, the default spread, the term spread, and the T-bill rates are statistically significant predictors in-sample.
5. Transition probabilities are a function of the One-month T-Bill rate, while the predictors mentioned are default premium, change in money stock and dividend yield.
6. Given the assumption of no contemporaneous effect, the GLS estimator is consistent and asymptotically efficient. Furthermore, the maximum likelihood estimator (MLE) and an equation-by-equation OLS estimator both yield the same result (see Hamilton, 1994).
7. The two sum of squares residuals statistics are a function of the selected parameters z^* and d.
8. Since regimes are exhaustive and mutually exclusive, such a specification does not suffer of dummy variable trap. In other words, $\sum_{k=1}^{K} D_{kt}(z) = 1$ for any t.
9. Given a continuous random variable $X \in \mathbb{R}$ and a Bernoulli random variable $D \sim B(1,p)$: $\mathrm{E}[DX] = \mathrm{E}[DX|D=1]p + \mathrm{E}[DX|D=0](1-p) = \mathrm{E}[X|D=1]p$.
10. The higher the number, the smaller the standard errors of the estimates $\hat{p}_{k(t+2)}$ for k = 1, 2, …, K.
11. For any given number n of simulations between each period, the total number of simulations needed to perform the iterated forecast is n^{h-1}.
12. The estimation is actually performed by *Quasi* Maximum Likelihood Estimation (QMLE), as the distribution of the errors is unknown. QMLE estimates are consistent if the conditional mean function and the conditional variance function are correctly specified (Bollerslev and Wooldridge, 1992).
13. In the following, the subscript indicates the iteration number, while the accent $\sim$ indicates that the estimate is still under an iteration algorithm.
14. Given two subsequent iteration having estimates $\boldsymbol{\theta}'$ and $\boldsymbol{\theta}$, convergence is achieved if $\boldsymbol{\theta}' - \boldsymbol{\theta} < \bar{\varepsilon}$, with $\bar{\varepsilon}$ being a small number (usually e^{-4}). In other words, when the two subsequent iterations leave, the estimated parameters approximately unaltered.
15. Autocorrelation stems from the transition equation in (7).
16. Kenneth French's data repository is freely accessible on the following website: http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.
17. We always refer to continuously compounded returns, i.e., log returns. Annualized returns are therefore calculated in the following way: $\bar{r}^{\text{annualized}} = 12\bar{r}^{\text{monthly}}$.
18. We calculate annualized volatility in the following way: $\sigma^{\text{annualized}} = \sqrt{12}\sigma^{\text{monthly}}$. Such computation is based on the assumption that returns IID. Although the underlying hypothesis is strong, the measure is still useful for the comparison of summary statistics.
19. In other words, when making a forecast from time t to time $t+h$, we only use data up to time t.
20. In order to perform strictly out-of-sample forecasts, the goodness of fit of the models should be tested at each iteration.
21. The analysis of forecast accuracy of MSI and MSIH models with up to 6 regimes does not lead to any different

outcome. In particular, their prediction errors rank very similarly to the corresponding models with 3 regimes.

22. The *saturation ratio* of a model is defined as the ratio between the number of observations available and the number of parameters to be estimated. Values of the saturation ratio between 15 and 20 are sometimes accepted in the literature. We set 20 as a threshold because, in performing the out-of-sample forecasts, we estimate models on expanding subsamples, in which a smaller number of observation. The earliest window of data in the analysis (1929:01–1979:12), for instance, includes 612 observations. A saturation ratio of 20 on the full sample is just 12.14 for this subsample.
23. Given a sample size n, an estimator $\hat{p}$ of the true specification of the model is *consistent* if: $\lim_{n\to\infty} Pr(\hat{p}_{\mathrm{sel}} = p_{\mathrm{true}}) = 1$, where $\hat{p}_{\mathrm{sel}}$ is the specification chosen by the estimator and p_{true} is the true, but unknown specification.
24. Alternatively, $\mathrm{E}_t\left(e^{j,\mathcal{M}_i}_{t+h}\right) = r^j_{t+h} - \mathrm{E}_t\left(r^{j,\mathcal{M}_i}_{t+h}\right)$.
25. The *bias* of an estimator $\hat{\theta}$ is defined as $B\left(\hat{\theta}\right) = \mathrm{E}\left(\hat{\theta}\right) - \theta$. An estimator $\hat{\theta}$ is *unbiased* if and only if its bias $B\left(\hat{\theta}\right)$ is 0, i.e., $\mathrm{E}\left(\hat{\theta}\right) = \theta$.
26. An identical approach involves the following regression: $\hat{e}^{j,\mathcal{M}_i}_{t,t+h} = \gamma^j_{0,h} + \gamma^j_{1,h}\hat{r}^{j,\mathcal{M}_i}_{t,t+h} + \xi^{j,\mathcal{M}_i}_{t,t+h}$. The forecast model is unbiased if $\gamma^j_{0,h} = 0$ and $\gamma^j_{1,h} = 0$.
27. This is similar to Guidolin and Timmermann (2008a, b) who acknowledge that the best specification for the forecasts on the Size and Value factors does not include autoregressive terms.
28. Forecast errors in a h-period forecast follow $MA(h-1)$ processes. Diebold and Mariano (1995) use the Newey–West estimator because it corrects for the resulting autocorrelation.
29. Hong and Lee (2003) study two additional loss functions: trading return and correct direction. They argue that squared and absolute loss may not be appropriate for financial returns, as investors eventually aim at maximizing profits, rather than forecast accuracy.
30. Under mean squared loss, if the dependent variable and its predictors are joint covariance stationary, optimal forecast errors have 0 mean (i.e., forecasts are unbiased), have variance which is non-decreasing function of the time horizon h, and are at most serially correlated of order $h-1$. Finally, single-period forecast errors are serially uncorrelated (see Diebold and Lopez, 1996).
31. We calculate the standard error of the mean as $\sigma/\sqrt{n}$, where σ is the standard deviation of the estimates and n is number of estimates.
32. In a multivariate framework, we model the joint probability density function of factor-mimicking portfolio excess returns. In a univariate framework, instead, we model the probability density function of each factor, independently. Regime-switching models are more flexible in the latter case, because regime shifts are tailored in order to capture the non-normalities that characterize each times series of excess returns.
33. The use of Markov switching forecast models does not pose an issue. With this regard, Buckley *et al*, (2008) consider portfolio optimization problems where asset returns are distributed as a mixture of multivariate normals.
34. For instance, market return dispersion (see Stivers and Sun, 2010) or other macroeconomic variables (see Sarwar *et al*, 2015).

REFERENCES

Ang, A. (2014) *Asset Management: A Systematic Approach to Factor Investing*. Oxford University Press.

Ang, A. and Timmermann, A. (2012). Regime changes and financial markets. *Annual Review of Financial Economics* 4: 313–337.

Ang, A., Goetzmann, W.N. and Schaefer, S.M. (2010). The efficient market theory and evidence: implications for active investment management. *Foundations and Trends in Finance* 5: 157–242.

Angelidis, T. and Tessaromatis, N. (2014) Global portfolio management under state dependent multiple risk premia. In: *Proceedings of Economics and Finance Conferences (No. 0400966)*. International Institute of Social and Economic Sciences.

Arnott, R., Kelso, C., Kiscadden. R. and Macedo, R. (1989) Forecasting factor returns: an intriguing possibility. *Journal of Portfolio Management* 16: 28–35.

Black, A.J. and McMillan, D.G. (2004). Non-linear predictability of value and growth stocks and economic activity. *Journal of Business Finance & Accounting* 31: 439–474.

Blitz, D.C. and van Vliet, P. (2007) The volatility effect. *The Journal of Portfolio Management* 34: 102–113.

Bollerslev, T. and Wooldridge, J.M. (1992). Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariances. *Econometric Reviews* 11: 143–172.

Buckley, I., Saunders, D. and Seco, L. (2008) Portfolio optimization when asset returns have the Gaussian mixture distribution. *European Journal of Operational Research* 185: 1434–1461.

Campbell, J.Y. and Cochrane, J.H. (2000). Explaining the poor performance of consumption-based asset pricing models. *Journal of Finance* 55: 2863–2878.

Campbell, J.Y. and Yogo, M. (2005) Implementing the econometric methods in "Efficient tests of stock return predictability". Unpublished working paper. University of Pennsylvania.

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

Chan, K.S. (1993) Consistency and limiting distribution of the least squares estimator of a threshold autoregressive model. *The Annals of Statistics* 21: 520–533.

Chen, L., Petkova, R. and Zhang, L. (2008). The expected value premium. *Journal of Financial Economics* 87: 269–280.

Chordia, T. and Shivakumar, L. (2002). Momentum, business cycle, and time-varying expected returns. *Journal of Finance* 57: 985–1019.

Cochrane, J.H. (2008) The dog that did not bark: A defense of return predictability. *Review of Financial Studies* 21: 1533–1575.

Cohen, R. B., Polk, C. and Vuolteenaho, T. (2003) The value spread. *Journal of Finance* 58:609–642.
Cooper, M.J., Gutierrez, R.C. and Hameed, A. (2004) Market states and momentum. *Journal of Finance* 59: 1345–1365.
DeMiguel, V., Garlappi, L. and Uppal, R. (2009). Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy? *Review of Financial Studies* 22(5): 1915–1953.
Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B* 39: 1–38.
Diebold, F.X. and Lopez, J.A. (1996) Forecast evaluation and combination. In: Maddala and Rao (eds.) *Handbook of Statistics*. Amsterdam: Elsevier.
Diebold, F. and Mariano, R. (1995) Computing predictive accuracy. *Journal of Business and Economic Statistics* 13: 253–263.
Doan, T., Litterman, R. and Sims, C. (1984) Forecasting and conditional projection using realistic prior distributions. *Econometric Reviews* 3: 1–100.
Fama, E.F. (1990) Stock returns, expected returns, and real activity. *Journal of Finance* 45(4): 1089–1108.
Fama, E.F. and French, K.R. (1988) Dividend yields and expected stock returns. *Journal of Financial Economics* 22(1): 3–25.
Fama, E.F. and French, K.R. (1993) Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* 33: 3–56.
Fama, E.F. and French K.R. (1995) Size and book-to-market factors in earnings and returns. *Journal of Finance* 50: 131–155.
Fama, E.F. and French, K.R. (2002) The equity premium. *Journal of Finance* 57(2): 637–659.
Fama, E.F. and French, K.R. (2010) Luck versus skill in the cross-section of mutual fund returns. *Journal of Finance* 65: 1915–1947.
Ferson, W., Siegel, A.F. and Xu, P.T. (2006) Mimicking portfolios with conditioning information. *Journal of Financial and Quantitative Analysis* 41: 607–635.
Giacomini, R. and White, H. (2006) Tests of conditional predictive ability. *Econometrica* 74: 1545–1578.
Guidolin, M. (2012). Modelling, estimating and forecasting financial data under regime (Markov) switching. Working paper, Bocconi University.
Guidolin, M. and Ono, S. (2006) Are the dynamic linkages between the macroeconomy and asset prices time-varying? *Journal of Economics and Business* 58: 480–518.
Guidolin, M. and Timmermann, A. (2006). An econometric model of nonlinear dynamics in the joint distribution of stock and bond returns. *Journal of Applied Econometrics* 21: 1–22.
Guidolin, M. and Timmermann, A. (2008) International asset allocation under regime switching, skew, and kurtosis preferences. *Review of Financial Studies* 21: 889–935.
Guidolin, M. and Timmermann, A. (2008) Size and value anomalies under regime shifts. *Journal of Financial Econometrics* 6: 1–48.
Guidolin, M., Hyde, S., McMillan, D. and Ono, S. (2009) Non-linear predictability in stock and bond returns: When and where is it exploitable? *International Journal of Forecasting* 25: 373–399.
Gulen, H., Xing, Y. and Zhang, L. (2011) Value versus growth: time-varying expected stock returns. *Financial Management* 40: 381–407.
Hamilton, J.D. (1990) Analysis of time series subject to changes in regime. *Journal of econometrics* 45: 39–70.
Hamilton, J.D. (1994). *Time series analysis*. Princeton: Princeton University Press.
Hansen, B.E. (2000). Sample splitting and threshold estimation. *Econometrica* 68: 575–603.
Haugen, R.A. and Baker, N.L. (1991) The efficient market inefficiency of capitalization-weighted stock portfolios. *Journal of Portfolio Management* 17: 35–40.
Hong, Y. and Lee, T.H. (2003) Inference on predictability of foreign exchange rates via generalized spectrum and nonlinear time series models. *Review of Economics and Statistics* 85: 1048–1062.
Ilmanen, A. (2011). *Expected Returns: An Investor's Guide to Harvesting Market Rewards*. John Wiley & Sons.
Inoue, A. and Kilian, L. (2005) In-sample or out-of-sample tests of predictability: which one should we use? *Econometric Reviews* 23: 371–402.
Kim, D., Roh, T.Y., Min, B.K. and Byun, S.J. (2014). Time-varying expected momentum profits. *Journal of Banking & Finance* 49: 191–215.
Krolzig H.-M. (1997) *Markov-Switching Vector Autoregressions: Modelling, Statistical Inference, and Applications to Business Cycle Analysis (Lecture Notes in Economics and Mathematical Systems)*, vol. 454. Berlin: Springer.
Leroux, B.G. (1992) Maximum-likelihood estimation for hidden Markov models. *Stochastic Processes and their Applications* 40: 127–143.
Lintner, J. (1965). The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics* 47: 13–37.
Lynch, A. (2000). Portfolio choice and equity characteristics: characterizing the hedging demands induced by return predictability. *Journal of Financial Economics* 62: 67–130.
Marcellino, M., Stock, J.H. and Watson, M.W. (2006). A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. *Journal of Econometrics* 135: 499–526.
Merton, R.C. (1973) An intertemporal capital asset pricing model. *Econometrica* 41: 867–887.
Mincer, J.A. and Zarnowitz, V. (1969) The evaluation of economic forecasts. In *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance* (pp. 3–46). NBER.
Panopoulou, E. and Plastira, S. (2014) Fama French factors and US stock return predictability. *Journal of Asset Management* 15: 110–128.
Perez-Quiros, G. and Timmermann, A. (2000) Firm size and cyclical variations in stock returns. *Journal of Finance* 55: 1229–1262.
Pesaran, M.H. and Timmermann, A. (1995). Predictability of stock returns: Robustness and economic significance. *The Journal of Finance* 50(4): 1201–1228.
Rapach, D.E. and Wohar, M.E. (2006). In-sample vs. out-of-sample tests of stock return predictability in the context of data mining. *Journal of Empirical Finance* 13: 231–247.
Rapach, D.E., Wohar, M.E. and Rangvid, J. (2005). Macro variables and international stock return predictability. *International Journal of Forecasting* 21: 137–166.

Roeder, K. and Wasserman, L. (1997). Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association* 92: 894–902.

Sarwar, G., Mateus, C. and Todorovic, N. (2015) Macroeconomic determinants of cyclical variations in value, size and momentum premiums in the UK. Size and Momentum Premiums in the UK, working paper, Cass Business School.

Sharpe, W. F. (1964) Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance* 19: 425–442.

Stivers, C. and Sun, L. (2010) Cross-sectional return dispersion and time variation in value and momentum premiums. *Journal of Financial and Quantitative Analysis* 45: 987–1014.

Tong, H. (1983) Threshold models in non-linear time series analysis. Lecture Notes in Statistics, No. 21.

Tsay, R.S. (1998) Testing and modeling multivariate threshold models. *Journal of the American Statistical Association* 93: 1188–1202.

van Dijk, D. and Franses, P.H. (2003) Selecting a nonlinear time series model using weighted tests of equal forecast accuracy. *Oxford Bulletin of Economics and Statistics* 65: 727–744.

Wang, K.Q. and Xu, J. (2015). Market volatility and momentum. *Journal of Empirical Finance* 30: 79–91.

Zakamulin, V. (2013) Forecasting the size premium over different time horizons. *Journal of Banking & Finance* 37: 1061–107.

Zakamulin, V. (2014) Predictable dynamics in the Small Stock Premium. *Economics Research International* 2014: 1–12.