1 Introduction

Factors in the cross-section of stock returns are persistent sources of risk premia (Fama and French (1993)), forming the basis for factor investing. Factors such as value and momentum deliver high returns over the long run, but can underperform in the short run, creating the opportunity for investors to engage in factor timing. Many investors use multi-factor portfolio strategies that tilt towards (buy) factors that are likely to outperform, and tilt away from (sell) factors that are likely to underperform, which raises several interesting questions. First, can we predict the probability of a factor outperforming or underperforming? Second, what is the economic significance of these predictions? Third, which variables are influential for factor timing, and do their nonlinear interactions matter? This paper is dedicated to answering these questions, consequently enabling a more reliable investigation into the economic mechanisms that drive factor risk premia.

There are four major challenges of factor timing that the literature so far has struggled to overcome in a unified framework. First, factors are a function of many financial and macroeconomic variables (Bender et al. (2018), Dong et al. (2022)). Second, the functional form of factors is unknown and likely complex, depending on nonlinear interactions among the variables (Didisheim et al. (2023), Gu et al. (2020), Kelly et al. (2023)). Third, factors are time-varying and depend on short- and long-term macroeconomic and financial conditions (Ilmanen et al. (2021), Hodges et al. (2017), Polk et al. (2020)). Fourth, factors share a common structure that should be summarized using a few latent features (Haddad et al. (2020), Kagkadis et al. (2023)). While the literature has largely used linear models that can address some of these challenges, none of these models, to our knowledge, can overcome all of them. As a result, the fundamental question concerning the feasibility of factor timing remains contentious (Asness (2016), Asness et al. (2017), Dichtl et al. (2019)).

In this paper, we introduce a deep learning approach for factor timing that addresses the aforementioned challenges by incorporating economically motivated restrictions. Since factor timing is fundamentally a prediction problem, employing neural networks is a natural

choice, as they excel at handling a large number of variables and complex functional forms. However, off-the-shelf neural networks are typically designed for prediction tasks in static environments with high signal-to-noise ratios and a large number of observations. However, factors exhibit time-varying risk premia, low signal-to-noise ratios, and limited historical observations, leading off-the-shelf neural networks to overfit. Hence, economic restrictions play a crucial role in regularization and are essential for improving the out-of-sample predictive accuracy of neural networks.

We demonstrate how to improve deep learning models for factor timing by incorporating economic structure. Our crucial innovation is to develop a multi-task neural network (MT) architecture to jointly predict all factors within a single functional form. MT learns common latent features across factors and uses factor-specific layers to capture each factor's nonlinear exposures to these latent features. This economic restriction offers two main benefits that improve the model's out-of-sample generalizability. First, MTs incorporate the economic restriction that all factors stem from a low-dimensional set of common latent features, which allows the model to capture commonalities among factors and improves performance. This shared representation has been shown to enhance data efficiency (Caruana, 1997), which is critical for applying deep learning to the limited historical observations of factors. In contrast, single-factor models are unable to leverage the low-dimensional common structure across factors, which makes them susceptible to overfitting on factor-specific noise. Second, multi-task learning (MTL) acts as a regularization technique that compels MT to generalize across multiple outputs (Ruder, 2017), thereby preventing overfitting to the noise of a single factor. This safeguard against overfitting is critical for factors, given their low signal-to-noise ratios.

Furthermore, we incorporate a high-dimensional set of variables and capture the time-variation of the factors as a function of macroeconomic and financial conditions using two separate recurrent Long Short-Term Memory neural networks (LSTMs), which perform dimension reduction and extract time series dynamics. The first LSTM takes in a large number

of macroeconomic time series as inputs, and summarizes their dynamics into a small number of macroeconomic state processes; while the second LSTM takes in a large number of financial time series as inputs, and summarizes their dynamics into a small number of financial state processes. Chen et al. (2023) demonstrate that LSTMs can capture short and long-term dependencies, which are necessary for detecting cycles, hence our LSTMs are designed to capture business and financial cycles. The economic restriction of separately modeling the macroeconomic and financial cycle dynamics reduces the number of parameters, and further enhances the neural network's out-of-sample performance. Augmenting the MT architecture with LSTMs gives rise to the dynamic multi-task neural network (DMT) model to address all four major challenges of factor timing.

We study six well-known factors over 57 years from January 1965 to December 2021. These factors include the excess market return (MKT), size (SMB), and value (HML) from Fama and French (1996), profitability (RMW) and investment (CMA) from Fama and French (2015), and momentum (MOM). Our predictors consist of 272 variables, including 123 macroeconomic variables from McCracken and Ng (2016), and 149 financial variables from Chen and Zimmermann (2022) and Welch and Goyal (2008). In each month, we condition on these predictors to forecast the probability of each factor earning a positive risk premium, i.e., the factor goes up. We compare MT and DMT, which predict the probabilities of all factors in a single functional form, against off-the-shelf models that estimate a separate functional form for each factor. These off-the-shelf classification models include logistic regression (LR), penalized logistic regression (EN), random forest (RF), extremely randomized trees (XRF), gradient boosted trees (GBT), support vector machine (SVM), and feed-forward neural network (NN).

Our results advance the knowledge on factor timing in five main dimensions. First, we assess the predictive power of the different models for factor timing using the out-of-sample predictive accuracy metric, defined as the proportion of correctly classified excess return directions in the out-of-sample period from January 1990 to December 2021. We show that

economic restrictions matter for the average accuracy across factors. Linear models, LR and EN, post accuracies of 53.9% and 54.1%, respectively, underperforming the 55.3% accuracy of the buy-and-hold benchmark (Buy) that always predicts a positive excess return. In contrast, nonlinear models based on decision trees RF, XRF, and GBT deliver higher accuracies of 56.8%, 55.9%, and 54.6%, respectively. SVM and NN exhibit the lowest accuracies of 53.5% and 52.3%, respectively, since they are too flexible and overfit to factor-specific noise. However, with the imposition of an economically motivated restriction of a common structure across factors, MT raises the accuracy to 55.7%, suggesting that the regularization and data efficiency benefits gained from MTL improve forecasting accuracy. Furthermore, by incorporating time series dynamics, DMT delivers the highest average accuracy of 57.2%. Notably, it is the only model to outperform the Buy for every factor, and is the most accurate model for MKT, RMW, CMA, and MOM.

We also conduct pairwise comparisons between different machine learning models using the Diebold and Mariano (1995) test statistic, and use the average log loss across factors to compare probability forecasts between models. DMT significantly outperforms all other models with t-statistics ranging from 2.97 to 13.27, underscoring the importance of incorporating economic structure and time series dynamics into factor timing models.

Second, we study whether factor timing using machine learning models can be exploited in an economically significant trading strategy. We employ a strategy that buys factors if the model predicts a positive return and invests in the risk-free rate otherwise. Our multi-factor strategy is then an equal-weighted portfolio of these strategy excess returns across all factors. We find that the economic significance of models aligns very closely with their average accuracies. Linear models LR and EN, with Sharpe ratios of 0.84 and 0.83 respectively, underperform compared to the multi-factor Buy Sharpe ratio of 0.98, attributable to their less accurate forecasts. In contrast, nonlinear models RF and XRF, with Sharpe ratios of 1.07 and 1.01, outperform both linear models and the multi-factor Buy, whereas GBT and SVM, each with a Sharpe ratio of 0.88, surpass the performance of linear models.

Turning to deep learning models, NN records the lowest Sharpe ratio of 0.78, since its lack of economic structure causes the model to overfit on factor-specific noise. In contrast, MT earns a Sharpe ratio of 1.14 by incorporating a common structure across factors, surpassing the leading off-the-shelf model (RF) Sharpe ratio of 1.07. Additionally, DMT offers the highest Sharpe ratio of 1.26 by further incorporating time series dynamics. DMT also achieves the highest alpha of 1.68% (t-stat of 4.29) with respect to the Buy, surpassing MT and RF's alphas of 0.98% (t-stat of 3.16) and 0.8% (t-stat of 2.38), respectively. Even after subtracting large transaction costs of 14 basis points, DMT earns an impressive Sharpe ratio of 1.17 and alpha of 1.33% (t-stat of 3.42). These results highlight the economic gains from incorporating economic structure and time series dynamics into deep learning models for factor timing.

Third, we dissect the multi-factor strategy by studying the economic significance of machine learning predictions for each individual factor. We find that DMT consistently outperforms, achieving a higher Sharpe ratio than the Buy for every factor. DMT is also the best performing model for MKT and CMA. Particularly for market timing, DMT achieves an alpha of 4.78% (t-stat of 4.17) and Sharpe ratio of 0.89, respectively, significantly outperforming MKT's Sharpe ratio of 0.6. Even after incorporating large transaction costs of 14 basis points, DMT's market timing strategy earns a Sharpe ratio of 0.87 and alpha of 4.5% (t-stat of 3.94). Additionally, we find that nonlinear models generally outperform linear models for the investment and momentum factors, suggesting that incorporating nonlinear interactions is important for understanding the economic mechanisms driving these two factors.

Fourth, we quantify the importance of different predictors for multi-factor timing using Shapley values, which approximate changes in the model probability predictions had we excluded certain predictors in its estimation. Models generally agree on the most important variables across factors. Influential financial variables include tail risk beta; price trends (return seasonality years 11 to 15, momentum based on ff3 residuals, industry return of big

firms); and accounting variables in the leverage (organizational capital, industry concentration, composite debt issuance), value (net payout yield, equity duration, book to market), and profitability (earnings consistency, dividend omission, earnings surprise of big firms) categories. Influential macroeconomic variables fall into the money (total reserves of depository institutions, real estate loans at all commercial banks, consumer motor vehicle loans outstanding), output (ip: nondurable consumer goods, ip: residential utilities), labor (civilians unemployed for 15-26 weeks, average duration of unemployment), and inflation (cpi: apparel, ppi: metals and metal products, ppi: finished goods, personal cons. exp: durable goods) categories.

Finally, we analyze DMT's most influential variables for each factor, revealing heterogeneity across factors. For MKT, Industrial Production (IP) related to nondurable consumer goods emerges as the most influential predictor. Additionally, we find that price trends, as well as variables from every accounting category, play a significant role in market timing. In contrast, for SMB, the most influential variables include tail risk beta and those in the leverage and liquidity categories, reflecting the key vulnerabilities of small firms. Accounting-based factors HML, RMW, and CMA possess a common set of influential variables, particularly those in the leverage, profitability, value, and investment categories. Finally, the influential variables for MOM predominantly include tail risk and price trends. However, we discover that variables in the value, leverage, profitability, and inflation categories also significantly influence the momentum factor. Notably, we find that many asset pricing anomalies provide significant predictive power for factor timing, in agreement with Dong et al. (2022) but contrasting the results of Cakici et al. (2023) and Engelberg et al. (2023). Furthermore, the substantial predictive power of financial variables is enhanced by models that account for nonlinear interactions. For example, we show pairwise nonlinear interactions for each factor, elucidating how the superior performance of DMT is attributable to its ability to capture these complex functional forms.

In this paper, we develop economically motivated deep learning models for factor timing,

contributing to at least two existing strands of literature. Firstly, we build upon the extensive literature that investigates factor predictability through linear models. Market predictability is a central topic in financial economics. Recent comprehensive reviews by Koijen and Van Nieuwerburgh (2011) and Rapach and Zhou (2013) underscore the depth and breadth of research in this domain. Various papers extend the ideas of market predictability to specific factors. For instance, the predictability of the value factor is explored by Baba Yara et al. (2021) and Cohen et al. (2003). Similarly, the predictability of the momentum factor is studied in Cooper et al. (2004) and Daniel and Moskowitz (2016).

Several recent studies delve into multi-factor predictability. Greenwood and Hanson (2012) employ corporate share issuance to predict multiple factors. Moreira and Muir (2017) adopt volatility approaches for multi-factor timing. Haddad et al. (2020) and Kagkadis et al. (2023) harness dimension reduction techniques for their factor predictions. Furthermore, Gupta and Kelly (2019) and Moskowitz et al. (2012) document factor persistence. A common theme among these studies is their reliance on linear models, a limited set of predictors, and a lack of time series dynamics. In contrast, we introduce the DMT model that adeptly captures time series dynamics, nonlinear interactions, a wide range of predictors, and commonalities across factors.

Second, we contribute to the rapidly expanding literature that uses machine learning techniques to forecast stock returns. In their seminal work, Gu et al. (2020) employ an extensive array of off-the-shelf machine learning techniques to forecast stock portfolio returns. We benchmark our analysis against their top-performing model based on a deep neural network with three hidden layers. Our findings demonstrate that incorporating a common structure across factors and time series dynamics yields superior predictions compared to all off-the-shelf approaches.

Several recent papers introduce machine learning models with economic restrictions to improve forecasts. Bryzgalova et al. (2023), Chen et al. (2023), Feng et al. (2018), Gu et al. (2021), and Kozak et al. (2020) develop machine learning techniques with the economic

restriction of no-arbitrage for predicting individual stock returns. Guijarro-Ordonez et al. (2021) develop a deep learning approach for statistical arbitrage. Liu (2023) develops multi-task neural networks that incorporate an economic restriction based on a common structure across return quantiles, which enhances the accuracy of quantile forecasts, resulting in significant statistical and economic gains. Proner (2023) introduces dynamic multi-task neural networks that jointly forecast inflation and unemployment, incorporating an economic restriction that is motivated by the Phillips Curve. Our paper complements this literature by developing a deep learning approach with economic restrictions, tailoring it specifically for the four main challenges of factor timing.

The rest of the paper is organized as follows. Section 2 lays out the model framework. We detail the data and estimation procedures in Section 3. Section 4 presents the main empirical results. Finally, Section 5 offers concluding remarks.

2 Methodology

We are interested in predicting a factor's probability of outperforming or underperforming. We use a classification approach for factor timing for two main reasons. First, the probability prediction naturally translates into an intuitive trading strategy. Second, factors exhibit heavy tails (Arnott et al. (2019), Daniel and Moskowitz (2016)), which can be detrimental for regression approaches, but has no effect on classification approaches.

In its most general form, we describe a factor's conditional probability of a positive excess return as

$$\pi_{i,t+1} \equiv P_t(r_{i,t+1} > 0) = g_i(x_t), \tag{1}$$

where factors are indexed by $i = 1, ..., N$, months by $t = 1, ..., T$, and the excess return of factor i in month $t + 1$ is denoted as $r_{i,t+1}$. Our objective is to estimate a functional form $g_i(\cdot)$, which links the P-dimensional vector of predictors x_t to the probability of a positive

excess return $\pi_{i,t+1}$. To achieve this objective, we introduce two multi-task neural network architectures that jointly model the $\pi_{i,t+1}$ across all N factors within a single functional form and incorporate several other economically motivated restrictions on g_i, which are detailed in the next sections.

2.1 Multi-task Neural Network

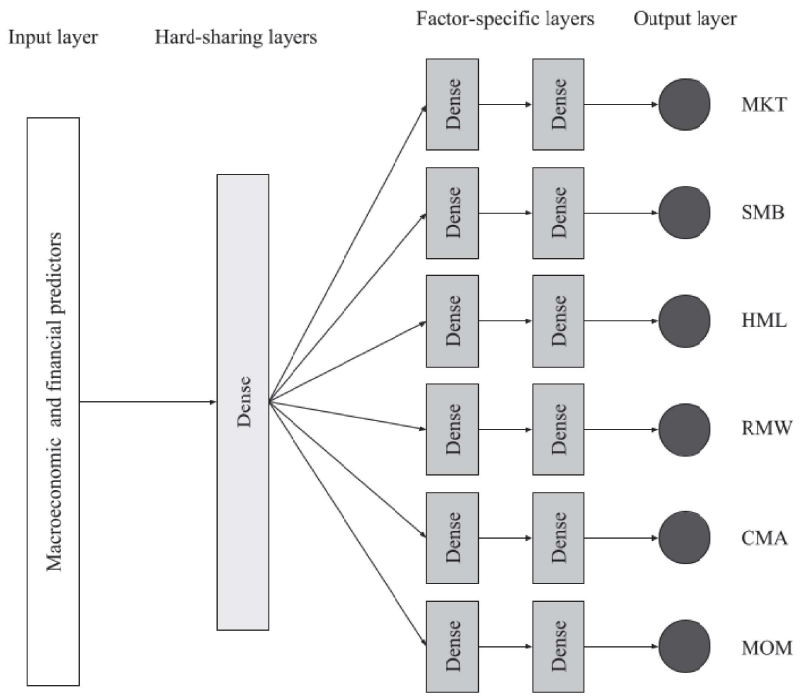
Factors share a low-dimensional common structure (Haddad et al. (2020), Kagkadis et al. (2023)). We incorporate this economic structure with a multi-task neural network architecture. Figure 1 shows an example of a MT. The "hard sharing" hidden layers interact and nonlinearly transform the input predictors into shared latent features, imposing an economically motivated restriction of a low-dimensional common structure across factors. The "factor-specific" layers capture nonlinear exposures to these shared latent features, subsequently aggregating these into a final prediction of $\pi_{i,t+1}$ for each factor.

MT employs MTL, which is a form of regularization that reduces overfitting to factor-specific noise, especially in low signal-to-noise environments. Simultaneous prediction of multiple factors forces the neural network's hard-sharing parameters to be sufficiently versatile, providing signals for most or all factors. This simultaneous learning of related tasks, known as inductive transfer, improves generalizability, as what is learned for each factor can help other factors be learned better (Ruder, 2017). Consequently, MTL increases the effective sample size, due to the extra information contained in the training signals of related tasks (Caruana, 1997).

2.2 Dynamic Multi-task Neural Network

Factors are time-varying and depend on short- and long-term macroeconomic and financial conditions (Ilmanen et al. (2021), Hodges et al. (2017), Polk et al. (2020)). MT and off-the-shelf models only incorporate predictor values in the preceding period, which is insufficient for the model to learn long-term business and financial cycle dynamics. To address this,

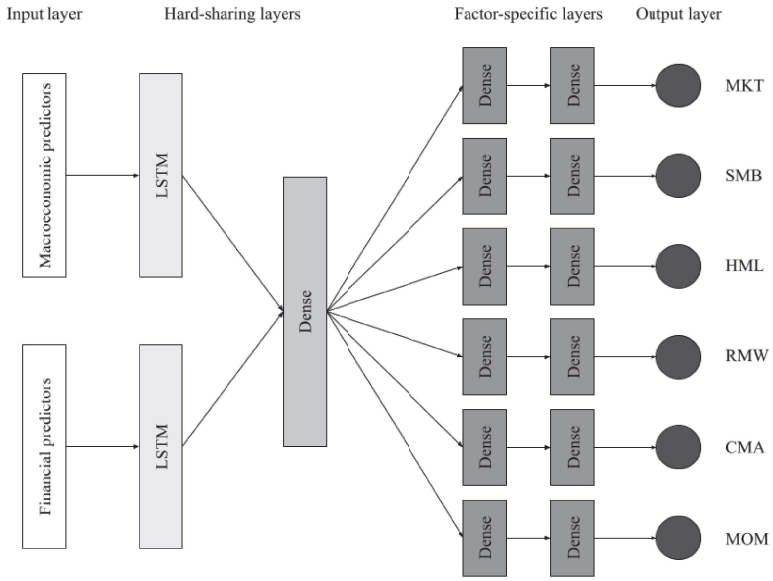
Figure 1: Multi-task Neural Network Example



Macroeconomic and financial predictors are input into a fully connected (dense) layer, where they are nonlinearly interacted. Next, the network splits into branches to learn factor-specific loadings and predict probabilities for each factor.

following Chen et al. (2023), we incorporate long-term nonlinear dependencies using LSTMs (Hochreiter and Schmidhuber, 1997). LSTMs allow for complex nonlinear interactions between the dynamics of the macroeconomic and financial time series, rather than just the values of the stationary time series at time t.

Figure 2: Dynamic Multi-task Neural Network Example



Macroeconomic and financial predictors are input into the network separately and propagate through LSTM layers constructing nonlinear dynamic features. These nonlinear dynamic macroeconomic and financial features are then nonlinearly interacted in a fully connected (dense) layer. Finally, the network splits into branches to learn factor-specific loadings and predict probabilities for each factor.

The DMT architecture, as illustrated in Figure 2, integrates memory by employing two separate LSTM layers for macroeconomic and financial variables, positioned before the first fully connected (dense) layer. The top LSTM nonlinearly interacts the macroeconomic time series inputs to estimate a small number of hidden macroeconomic state variables. Likewise, the bottom LSTM nonlinearly interacts the financial time series inputs to estimate a small number of hidden financial state variables. Drawing on Chen et al. (2023), these LSTM layers extract business and financial cycle dynamics through their ability to identify cyclical

patterns. The need to independently capture business and financial cycle dynamics drives our economic restriction to separate macroeconomic and financial input variables in the LSTM layer. A subsequent hard sharing layer nonlinearly interacts the high-level macroeconomic and financial hidden state variables, producing a low-dimensional set of shared latent features for the multi-task network.

Given the high-dimensionality and strong cross-sectional dependence of financial and macroeconomic input variables, there's overlapping information that can be captured by a low-dimensional model. Functionally, our LSTM layers serve to reduce dimensions, akin to principal component analysis (PCA), while also extracting dynamics akin to a state space model within a broader nonlinear framework. By consolidating the high-dimensional inputs into a small number of hidden state processes, we prevent the model from overfitting on the noise specific to each macroeconomic or financial time series. Internet Appendix IA1.2.5 provides a detailed description of the LSTM architecture.

2.3 Off-the-shelf Models

We consider a variety of off-the-shelf machine learning models from Hastie et al. (2009) as comparative benchmarks. These models estimate a separate functional form g_i for each factor i and can be highly susceptible of overfitting to factor-specific noise. Additionally, they do not leverage the common structure across factors. Off-the-shelf models studied in this paper include linear models (logistic regression and penalized logistic regression) and nonlinear models (random forest, extremely randomized trees, gradient boosted trees, support vector machines, and a feed-forward neural network with three hidden layers). These models are described in detail in Internet Appendix IA1.

3 Data and Estimation Procedure

3.1 Data

Our factor sample spans January 1965 to December 2021, totaling 57 years. We use six monthly factors from Kenneth French's website as response variables: MKT, SMB, HML, RMW, CMA, and MOM. The first three factors correspond to the excess market return, size, and value from Fama and French's original three-factor model (Fama and French, 1996). The following two factors indicate profitability and investment from their five-factor model (Fama and French, 2015), while the last factor represents momentum, which goes long on past winners and short on past losers.

Additionally, we include macroeconomic variables from the FRED-MD database, as detailed in McCracken and Ng (2016), which correspond to categories such as labor, output, and inflation. We apply their transformations to generate stationary time series. After removing variables with missing values before 1990, corresponding to the beginning of the out-of-sample period, we have 123 macroeconomic predictors. Following Bianchi et al. (2021) and Chen et al. (2023), we lag the macroeconomic variables by one additional month (i.e., we use the observation at $t-1$) to account for announcement delays. See Table IA3 in the Internet Appendix for a comprehensive list of macroeconomic variables.

We obtain 149 financial variables from two sources. First, motivated by Dong et al. (2022), we include "anomalies" from the asset pricing literature as predictors, specifically referring to long-short portfolios of individual stocks. We include the anomalies detailed in Chen and Zimmermann (2022), and categorize variables loosely based on Gu et al. (2020) and Jensen et al. (2023), encompassing investment, liquidity, price trend, profitability, quality, risk, and value.[1] We remove predictors with any missing values before 1990, corresponding to the beginning of the out-of-sample period, resulting in 137 predictors, and fill the remaining missing values with the expanding training set mean. Second, we include 12 aggregate

[1]The monthly data are available at www.openassetpricing.com. Our data is collected from the August 2023 Release.

predictors from Welch and Goyal (2008) based on stock and bond markets. Stock market variables include the book-to-market ratio, dividend price ratio, dividend yield, earnings price ratio, dividend payout ratio, stock variance, and net equity expansion; and bond market variables include the t-bill rate, long term yield, long term rate of return, term spread, default yield spread.[2] See Table IA2 in the Internet Appendix for a descriptive list of the financial variables.

3.2 Estimation Procedure

To estimate the models, we use the standard validation set approach and divide our full sample (January 1965 to December 2021) into three disjoint time periods. We start by estimating the model parameters on a training sample of 20 years (1965 - 1984). We then perform an extensive hyperparameter optimization, validating the model's fit in the next five years (1985 - 1989). Lastly, we assess the predictive power in the one-year testing sample (1990). We keep the models fixed for one year and replicate this procedure extending the number of years in the training sample by one year in each iteration, for a total of 32 out-of-sample years (1990 - 2021). In the training, validation, and test sets, we standardize each predictor using its mean and variance from the training set. Section IA3 in the Internet Appendix details the hyperparameter optimization setup.

4 Empirical Results

4.1 Time-varying Factor Risk Premia

Figure 3 displays the cumulative returns of the six factors across the entire sample. All factors gain considerable risk premia, with cumulative returns exceeding 100%. SMB fluctuates with periods of high and low returns, resulting in the lowest cumulative return. HML fares well before the financial crisis, but a significant post-crisis drawdown lands it with the second

[2]The monthly data are available at Amit Goyal's website.