UC Berkeley, Haas School of Business, Master of Financial Engineering, (2024)

THREE'S A CROWD OR THREE'S A PARTY? A JOINT CONSIDERATION OF REGIMES, FACTOR COMPRESSION, AND FACTOR ENGINEERING

YVES D'HONDT, MATTEO DI VENTI, AKSHITA GULATI, ROHAN RISHI AND JACKSON WALKER

Abstract. The key motivation behind this paper is to provide academics and professionals with an alternative to opaque, but powerful statistical and ML-based factor models. To achieve this, we present a structured factor model that combines classical approaches to factor modeling with novel research ideas. This model relies on three core ideas: regime-conditioning, factor selection, and factor engineering. The structured nature of our proposed model allows users to introspect the model end-to-end, gaining detailed insights into its inner workings. Over a cross-section of large-cap, US equities, we find moderate evidence on the benefits of our proposed model from its forecasting ability, and positive evidence from its ability to construct optimized portfolios.

§1. Introduction

This Applied Finance Project originated from an industry project on regime detection and dynamic portfolio construction [11], which provided evidence that regime-dependent asset allocation leads to portfolios with better risk-return characteristics. Looking towards the asset management industry, large factor models have been the status quo for many years in dealing with portfolio construction. These models come with unique challenges, most notably how to select relevant factors from an ever-increasing universe of factors, known as the *factor zoo*, and how to find new factors orthogonal to this factor zoo.

The key motivation behind this paper is to provide an alternative to opaque, but powerful statistical and ML-based factor models. To that end, we propose a structured and interpretable factor model. While we find moderate to strong evidence in favor of our proposed model, the model also represents a general framework that can be used and extended by academics and professionals to improve existing factor models and test the marginal usefulness of a new technique, in light of other available techniques.

We aim to expand the research around regime-dependent portfolio construction and factor models in multiple ways. First, whereas most papers have approached regime-dependent portfolio construction as an asset or index allocation problem, this paper applies and evaluates models over a large equity cross-section. Second, this paper considers the joint problem of regime-modeling, factor zoo compression, and finding new factors. Although each of these has been researched independently, to our knowledge they have not yet been considered jointly. Finally, this paper proposes a structured and explainable framework as an

Keywords: market regimes, factor models, portfolio optimization, optimal transport, HMM, GRS, SHAP, quantitative finance

alternative to recent developments in powerful, but opaque ML-based latent factor models.

This paper is structured as follows: Section 1 introduces the research project, delineating its scope and objectives. Section 2 provides readers with an overview of historical and recent developments surrounding factor modeling and regime detection. Section 3 outlines the key concepts and models that represent the building blocks of our proposed factor model. Section 4 follows up by detailing the structure and implementation of our proposed factor model. Section 5 gives a brief description of the evaluation data and Section 6 outlines the high-level evaluation framework. Section 7 presents the evaluation results and highlights the key insights. Before concluding, Section 8 presents a use case of the interpretable nature of our proposed factor model. Finally, Section 9 discusses the key findings and extensions to this project, while Section 10 summarizes the key findings and learnings.

§2. Background Knowledge & Recent Developments

Although some familiarity with factor models, regime detection, and portfolio optimization is helpful, this literature review provides all the necessary background knowledge for this paper. This section focuses on the history, current narrative, and recent research surrounding dynamic factor models. Technical details on the techniques used within this paper are outlined in Sections 3 and 4, which discuss the key concepts and implementation of the proposed factor model.

2.1. Factor Models

In its basic form, a factor model is a static, linear model that attempts to explain the expected return and (systematic) risk of one or more assets. This model can subsequently be used for various tasks ranging from ex-post performance analysis to ex-ante portfolio optimization. To enable this, the factor model offers two key outputs, namely an estimated asset variance-covariance matrix, referred to as the *VCV*, and an estimated asset expected excess return vector. Given N assets and L factors, the classical factor model is entirely defined by five matrices:

1. $\mathbf{F}$: an $L \times L$ factor covariance matrix.
2. $\mathbf{K}$: an $L \times N$ matrix with the factor loadings for each asset. This matrix is typically estimated through OLS.
3. $\mathbf{D}$: an $N \times N$ matrix with the residual or idiosyncratic covariance of the assets. This matrix is often diagonal in which case it contains the idiosyncratic variance of the assets.
4. $\mathbf{f}$: an $L \times 1$ vector with the expected returns for each factor.
5. $\boldsymbol{\alpha}$: an $N \times 1$ vector with the alpha for each asset.

Combining these five matrices, the asset VCV ($\hat{\Sigma}$) and expected excess return vector ($\hat{\mu}^e$) can be estimated as follows:

$$\hat{\Sigma} = K^T \cdot F \cdot K + D$$

$$\hat{\mu}^e = \alpha + K^T \cdot f$$

Following the pioneering results of Fama and French [14], the literature has seen a proliferation of proposed factors, giving rise to the term *factor zoo* to describe this phenomenon.

Although factor models were originally proposed to solve many of the numerical issues that arise from using empirical VCVs over large asset cross-sections, similar issues occur for large factor VCVs. To address this problem, Swade, et al. [39] compress the factor zoo by focusing on explaining the available alpha rather than the covariance matrix of factor returns. This compression exercise shows that in the US, only 15 factors are necessary to span the 153 factors considered [39].

In contrast to recent research on compressing the factor zoo, the never-ending search for new factors persists. Latent factor models are a popular tool for this with many innovations both on statistical models, such as RP-PCA [31], and ML models, such as Factor VAE [13] and HireVAE [43]. These models have shown good performance on paper, but the resulting factors are often opaque and non-intuitive, presenting a barrier to using them in the industry.

2.2. Regime Detection

Below we present a summary of the key literature surrounding regime detection. For a more in-depth discussion, we refer to our previous industry project on regime-dependent portfolio construction [11].

Tu provides evidence that there are losses associated with ignoring regime switching and that accounting for regime switching is substantially independent from incorporating model and parameter uncertainty in portfolio decisions [41]. Therefore, Tu argues that "the more realistic regime switching model is fundamentally different from the commonly used single-state model, and hence should be employed instead in portfolio decisions irrespective of concerns about model or parameter uncertainty" [41]. In line with Tu's observations, there have since been numerous papers that argue for incorporating regime dependency in investment decisions.

A key assumption behind regime detection is that markets are characterized by a number of latent, or hidden, regimes, each with their own return generating process. At any point in time returns are assumed to be generated from one of these regimes. Regime models come in many flavors, with each model having its own drawbacks and benefits. Broadly speaking, we group regime models into two categories: parametric and non-parametric models.

Hidden Markov Models (HMM) are a popular parametric model to detect latent regimes and have been studied in numerous financial applications [19][20][21][42]. At a high level, an HMM consists of a finite number of states with a fixed set of transition probabilities from one period to the next between each state. Next to this, each state is associated with its own returns-generating process, typically represented as a returns-distribution. For a more detailed explanation on the inner workings of HMMs, please refer to Section 3. Within the context of dynamic portfolio construction, Bae, et al. [1] and Costa, et al. [8][9] provide evidence in favor of using HMMs for regime detection.

Advances in machine learning have also led researchers to start investigating non-parametric models in latent regime detection. Bilokon, et al. [3] suggest using path signatures on returns time series in combination with a modified K-Means algorithm to

detect regimes. In similar fashion, Horvath, et al. [24] suggest a modified K-means algorithm using Wasserstein distance and Wasserstein Barycenters to detect regimes[1]. Although these papers offer promising initial results, as evidenced by our ability to construct improved portfolios using Wasserstein K-Means in our industry project [11], these methods are quite data-intensive and may be more suitable for higher-frequency analyses with more available data.

2.3. Dynamic Portfolio Construction

Regime detection adds a new dimension of understanding to market behavior, but it does not directly offer actionable recommendations. That is where regime-dependent portfolio construction comes into play. The idea behind regime-dependent portfolio construction is to identify and maximize objectives defined over a regime-switching model of asset returns. Based on the results of our previous industry project [11], this paper focuses on regime detection through HMMs.

An important feature of HMMs is that they provide transition probabilities between different states over time, given their interpretation as a state-transition model. This allows for stochastic portfolio optimization as regime sequences can be simulated from the fitted Markov chain. Bae, et al. use stochastic programming to construct optimal portfolios by maximising a portfolio objective over these sequences. They conclude that "the regime information helps portfolios avoid risk during left-tail events" [1]. One major drawback of stochastic programming is that it suffers from the *curse of dimensionality* and is therefore time and resource intensive when applied to a broad universe of assets or many states.

To overcome this drawback, Costa, et al. [8] suggest to use a simpler one-period model. They propose a regime-switching factor model, based on the Fama-French 3 Factor Model, that fully characterizes the systematic portion of the expected returns and covariance matrix of an asset universe at each point in time. They provide evidence that a risk-parity strategy based on this model offers higher returns at a similar ex-post level of risk compared to its nominal counterpart [8]. In a follow-up paper, Costa, et al. [9] propose a regime-switching factor model that allows for both systematic and idiosyncratic regime-dependency. They show that mean-variance optimization (MVO) using this novel framework consistently displays higher returns and similar or lower volatility than its nominal counterpart [9].

Another interesting recent development is the one put forward by Garleanu and Pedersen [16] that introduce factors models trying to optimize the dynamic trading of a portfolio. Under transaction costs and differing speed of mean reversion for signals, the optimal solution is to blend the current optimal portfolio with the optimal portfolio at next time steps.

[1] Wasserstein distance and Barycenters are concepts from optimal control theory. They provide a mathematically sound framework for defining a metric space between uni- or multivariate distributions, as well as a method for aggregating distributions within this space.

§3. Key Concepts & Models

As mentioned in the introduction, we consider the joint problem of regime modeling, factor selection, and factor engineering. This section introduces the key concepts, models, and papers on these individual topics. Subsequently, Section 4 builds on this section by detailing how each of these individual components is combined into a structured, dynamic factor model.

3.1. Selected Regime Model

Based on the results from our previous industry project [11], this paper employs Gaussian Mixture Model based Hidden Markov Model (GMMHMM) for regime detection. GMMHMM is relatively robust to overfitting, remains relatively stable during refitting, and is able to fit a wide range of potentially non-normal distributions, setting it apart from other regime-models. Furthermore, the empirical results from this industry project [11] showcase the superiority of GMMHMM within the context of regime-dependent portfolio optimization.

In general, Hidden Markov Models are a tool to model and forecast time-series data generated from a number of undetectable, or *latent*, states. Each state generates data from a state-dependent distribution. Moreover, the model incorporates a static *transition matrix* which describes the probabilities of a state change from one period to the next. A data generation process for each of the latent states together with a transition matrix defines an entire HMM model. GMMHMMs are a special case of HMMs which assume that the state-dependent distributions are represented by Gaussian Mixture Models. Appendix A offers an in-depth overview of the properties and assumptions behind HMMs.

This paper used `hmmlearn` to implement GMMHMMs. This package allows for control over two crucial hyperparameters: the *covariance type* of each regime's return distribution and the *emission model*. For the covariance type, this paper considers *diagonal* matrices, in which every state's covariance matrix is diagonal. For the emission model, this paper considers *Gaussian Mixture* emissions.

3.2. Factor Selection

As modern factor models grow in size, the curse of dimensionality can lead to an increase in numerical instability. Ideally, a factor model should have fewer factors than the number of assets in the portfolio, a goal that becomes challenging with models containing 100 or more factors. This paper partially builds on the work from Swade, et al. [39] and de Prado [10] who propose solutions to deal with high dimensional factor and asset cross-sections, respectively. This subsection introduces the relevant ideas from these papers, whereas Section 4 details their integration into our proposed factor model.

3.2.1. GRS for Factor Selection

The GRS test was originally proposed by Gibbons, Ross, and Shanken in 1989 as a way to test the ex-ante efficiency of a given portfolio [17]. Since then, the test has become a popular tool in asset pricing research due to its clearly defined test statistic distribution and its interpretation as a function of two Sharpe ratios. Nowadays, the GRS test is mainly

used to test the efficiency or rank the power of multiple factor models against each other.

Swade, et al. use the GRS test in their research to sequentially select factors in a static factor model [39]. We expand on their research by applying a similar method to our regime-dependent factor model. Despite the popularity of the GRS test, the original authors were ambiguous in how the test statistic should be constructed for a multivariate setting, such as to test multivariate factor models. Kamstra, et al. provide clarity with a detailed note and proof of the correct GRS test statistic [28]. Unfortunately, many researchers still use a wrong formulation of the GRS statistic, including Swade, et al. [39]. According to Kamstra, et al. even though the mistake appears minor, it can lead to over-rejection of factor models and more importantly misranking between factor models [28]. Appendix B provides the correct GRS specification used within this paper as per Kamstra, et al. [28].

3.2.2. Hierarchical Clustering of Financial Time Series

Most portfolio optimization techniques rely on Markowitz's mean-variance optimization (MVO), quadratic programming, and the estimation of an asset VCV. Although popular, these techniques can suffer from instability, concentration, and underperformance. de Prado proposes an alternative portfolio optimization technique called Hierarchical Risk Parity (HRP) which exploits the network structure of asset returns to overcome the aforementioned issues [10].

This paper draws inspiration from the network representation of asset returns. Given a universe of assets U of size N, we can think of each asset $a_i \in U$ as a node in a network. The network is fully connected as each asset a_i is potentially related to each other asset a_j. We can now think of the covariance between every two assets as the edge-relationship between those two assets. Given that a VCV matrix is symmetrical, it requires the estimation of $\frac{N(N+1)}{2}$ elements to represent the full network. For large N, this quickly becomes numerically unstable.

de Prado instead proposes to compress this network structure by performing hierarchical, agglomerative clustering over a custom asset-distance metric. This distance metric jointly considers the correlation between each asset and all other assets. The dendrogram resulting from this clustering finally offers an $O(N)$ representation of the asset-network versus the $O(N^2)$ representation required by an empirical VCV. Appendix C offers a detailed overview of the asset-distance metric.

de Prado goes on to show how this hierarchical representation of assets can be used directly for portfolio optimization through the HRP algorithm [10]. Subsequent research has also investigated how different portfolio objectives can be incorporated into this framework, such as the HERC algorithm from Raffinot [36], as well as how constraints can be added to this framework, such as the constrained HRP from Pfitzinger, et al. [34].

3.3. Factor Engineering

While a subset of factors may explain the equity cross-section in each regime, relying solely on unconditional factors may fail to capture the full scope of the equity cross-section.

This paper partially builds on the work from Ilic, et al. [25] and Lettau, et al. [31] who both offer algorithms that can be used or modified for factor engineering. This subsection introduces the relevant ideas from these papers, whereas Section 4 details their concrete implementation within this paper's framework.

3.3.1. Linear Boosting

Linear Boosting, a shorthand for Explainable Boosted Linear Regression (EBLR), is an iterative two-stage framework for time-series forecasting and non-linear feature generation proposed by Ilic, et al. [25]. In the first stage, a base learner[2] is fitted on the data. In the second stage, a single decision tree is fitted over the residuals from this base learner. The decision path leading to the leaf node that explains the largest portion of the residuals is subsequently encoded as a new binary feature for the base learner. Linear boosting alternates between these two stages, adding new features until some stopping criterion is reached.

EBLR draws upon the fact that regression trees aim to minimize MSE of the target values in terminal nodes, essentially grouping errors from the same source. Effectively, this potentially discovers non-linear features which explain a proportion of the errors. There are multiple hyperparameters in EBLR. For the purpose of this paper, the most important ones are the number of selected features, the three depth, and the minimum number of observations in any leaf node. These hyperparameters are crucial in preventing overfitting as they define the complexity of the generated non-linear features.

The advantage of EBLR over other models is twofold. First, since the non-linear features are generated from shallow decision trees, they are explainable and intuitive to understand. Second, EBLR offers a way to incorporate non-linearities in a purely linear model through feature encoding. As a result standard factor models can simply treat these non-linear features as any other factor time-series.

3.3.2. RP-PCA

RP-PCA draws from the intuition that standard PCA is not effective at identifying factors with small variance that still explain the expected returns of the assets well. To address this issue, Lettau and Pelger [31] develop a penalized PCA method that introduces an error term for mispricings of the expected returns. The fundamental intuition is that in asset pricing applications the first moment of the factors is equally as important as the second moment.

For N assets over T observations with returns $X \in \mathbb{R}^{T\times N}$, RP-PCA finds a set of factors F and factor loadings Λ such that the following objective is minimized:

$$\min_{\Lambda,F} \underbrace{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(X_{ti} - F_t\Lambda_i^\top\right)^2}_{\text{unexplained variation}} + \gamma \underbrace{\frac{1}{N}\sum_{i=1}^{N}\left(\bar{X}_i - \bar{F}\Lambda_i^\top\right)^2}_{\text{pricing error}} \tag{3.1}$$

[2] In general, any base learner can be used, including non-linear models.

Note that γ in Eq. (3.1) is a hyperparameter that determines the importance of the pricing error[3]. In the special case of $\gamma = -1$, RP-PCA is equivalent to standard PCA. According to Lettau and Pelger, the parameter is best set to higher values to strengthen the signal of the weak factors and at a comparatively smaller loss of efficiency. More generally, γ could be treated as any other hyperparameter in a model pipeline and be optimized for the objective at hand.

3.4. Factor Blending

Fitting a factor model conditional on a specific regime is as simple as fitting it over the subset of data that falls in that regime. This however leaves the question of how to blend each regime's factor model back into a single model. Costa, et al., propose a closed form solution to this problem [9]. They consider a two-state model, while their technique can easily be extended to allow for any finite number of regimes. First, the regime-dependent factor model is defined as follows for N assets, L factors, and 2 states:

$$r_t^e = I_{t,1}(\alpha_1 + K_1^T f_{t,1} + \epsilon_{t,1}) + I_{2,t}(\alpha_2 + K_2^T f_{t,2} + \epsilon_{t,2}) \tag{3.2}$$

Here $r_t^e \in \mathbb{R}^N$ is the asset excess return vector at time t, $I_{t,i}$ is an indicator function which is 1 if time t is in state i and 0 otherwise, α_i is the vector of fitted intercepts conditional on state i, $K_i \in \mathbb{R}^{L\times N}$ are the fitted factor loadings conditional on state i, $f_{t,i} \in \mathbb{R}^L$ are the factor returns at time t conditional on state i, and finally $\epsilon_{t,i} \in \mathbb{R}^N$ are the regression residuals at time t conditional on state i. By utilizing indicator functions for the regimes, this entire regime-dependent factor model can be fitted through OLS [9].

Costa, et al. then go on to show that under the assumption of normality of the regime-conditional factor returns and regression residuals, there is a closed form solution to arrive at a unique estimated expected return vector and asset VCV conditional on today's regime [9]. Let $\gamma_{i,j}$ be the probability of moving from state i to state j from this period to the next. Furthermore, assume that $\epsilon_{t,i} \sim N(0, D_i)$ and $f_{t,i} \sim N(0, F_i)$. D_i is suggested to be estimated as the diagonal-matrix of the VCV of the observed regression residuals for state i, while F_i is suggested to be estimated as the full empirical VCV of the factor returns for state i. Note that Costa, et al. implicitly assume that factor returns are mean-zero for each regime and thus might require a demeaning transformation in practice [9]. Given these assumptions, the asset expected excess return vector and VCV conditional on being in state i today are given by:

$$\hat{\mu}_i^e = \gamma_{i,1}\alpha_1 + \gamma_{i,2}\alpha_2$$

$$\begin{aligned}\hat{\Sigma}_i = {} & \gamma_{i,1}(K_1^T F_1 K_1 + D_1) + \gamma_{i,2}(K_2^T F_2 K_2 + D_2) \\ & + \gamma_{i,1}(1-\gamma_{i,1})\alpha_1\alpha_1^T + \gamma_{i,2}(1-\gamma_{i,2})\alpha_2\alpha_2^T \\ & - \gamma_{i,1}\gamma_{i,2}(\alpha_1\alpha_2^T + \alpha_2\alpha_1^T)\end{aligned}$$

[3] Effectively, RP-PCA applies SVD decomposition to find eigenvalues and eigenvectors of the modified covariance matrix: $\frac{1}{T}X^T X + \gamma \bar{X}\bar{X}^T$.

§4. Proposed Factor Model

Is there a benefit to the joint consideration of regime-conditioning, factor compression, and latent factor discovery? Equipped with the concepts from Section 3, we attempt to answer this question through the proposal of a structured factor model pipeline. Key to our proposed model is to maintain a high level of interpretability and explainability in the estimation process.

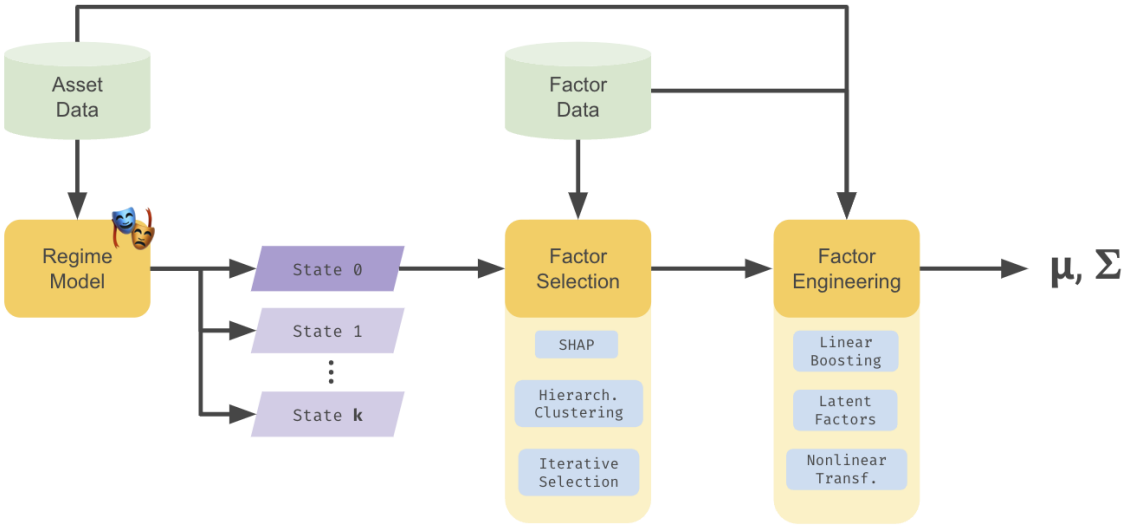

Figure 1. : High level structure of the proposed dynamic factor model. This model consists of three key components, namely: regime detection, factor selection, and factor engineering.

The high level structure of our proposed factor model pipeline can be seen in Figure 1. Following the methodology suggested by Bae, et al. [1] and Costa, et al. [8][9], our model begins by dividing historical data into K distinct regimes. Subsequently, for each regime, we independently apply factor selection and engineering. This way, the model implicitly allows for the idea that different factors could be important at different points in time, as alluded to by Swade, et al. [39]. After this step, a linear factor model is fitted for each regime, using the selected and engineered factors. Finally, each of the regime-dependent factor models is blended together to arrive at a single estimate for the asset VCV and expected excess return vector. Implemented as a unified pipeline, this model requires three synchronized inputs based on date indices: a time-series to fit regimes, a time-series panel of asset returns, and a time-series panel of factor returns.

In the following sub-sections, we outline the regime model, factor selection, and factor engineering techniques that we consider within this framework. Each of these three components is constructed in isolation such that we can plug and play with different combinations. Finally, we also outline how we blend the factor models for each regime back into a single estimate for the asset VCV and expected excess return vector. At the end of this section,

we also provide a brief overview of each component which can serve as a reference for the remainder of this paper.

4.1. Regime Model

As mentioned in Section 3.1, GMMHMM is used as the regime model in the proposed factor model. We fit this model using the daily market factor, Mkt-RF, as described by Fama-French [15]. Mkt-RF was chosen for its neutrality compared to specific indices such as the S&P500. This GMMHMM is automatically re-fitted whenever the factor model is fitted as the Baum-Welch algorithm uses both forward and backward passes through the data and hence a single fit over the full backtest period would constitute a significant lookahead bias.

In theory, an HMM can fit an arbitrary number of states, but real-world constraints limit the number of states. Most importantly, to ensure that the fitted factor VCV is positive semi-definite, we require $T >> L$ where T is the number of time-steps in the training sample and L is the number of factors. In cases where HMMs have numerous states, some states might be too sparse, resulting in $T < L$. This can cause degenerate factor VCVs in these states. Next to this, both the factor selection and engineering procedures can be data hungry, further requiring that each state has sufficient data. Moreover, a low number of regimes is often sufficient to show improved performance [11]. As a result, this paper only considers 2 and 3 state systems.

4.2. Factor Selection

Our proposed factor model builds on the idea that different market conditions may require a distinct number and combination of factors to effectively explain the equity cross-section. Within this study, we look at three distinct methods to select the relevant factors in each regime. First, in line with Swade, et al. [39], we develop a sequential factor selection method that selects one factor at the time, using the GRS test to decide on the most appropriate factor at each iteration. Next, we develop an agglomerative hierarchical clustering over factor returns using the correlation-based distance metric defined by de Prado as part of Hierarchical Risk Parity [10]. Finally, we develop a factor selection method that selects factors based on their feature importance in a large factor model measured through Shapley values.

In the following three subsections, we expand on the technical details of each of these three methods and the objective that they try to achieve.

4.2.1. Sequential Factor Selection With GRS

We slightly modify the factor selection procedure proposed by Swade, et al. [39] (see Section 3.2.1). Most notably, we apply the correct GRS specification and operate under a simplified stopping criterion. On a high level, given a large set of factors, the procedure attempts to find the smallest subset of factors that explains all the alpha of the other factors in the set. Concretely, the procedure works as follows:

1. Initialize the procedure by setting the CAPM as our current factor model, CFM.
2. For each remaining factor, consider a new factor model consisting of the CFM plus this new factor.

3. Compute the GRS statistics for each of those new factor models, using all factors not included in the new model as the test assets.
4. Rank the new factor models based on their GRS statistic and select the model with the lowest statistic.
5. Replace the CFM by this selected model.
6. Check the stopping criterion. If it is satisfied, return the CFM as the final model, else go back to step 2 and repeat until completion.

Different stopping criteria can be employed. In this paper, we adopt a straightforward criterion: the process stops when a predetermined number of factors have been selected.

4.2.2. Hierarchical Factor Selection

To retain the flexibility of quadratic and constrained optimization, we do not employ HRP. Rather, we use the agglomerative clustering proposed by de Prado [10] (see Appendix C) as a tool to reduce a large universe of factors into a smaller subset of representative factors. Concretely, our hierarchical factor selection procedure works as follows:

1. Given a large universe of factors, F, apply agglomerative clustering over $\tilde{D}(F)$[4] using Ward linkage.
2. Use the resulting dendrogram to return the K coarsest clusters from the dendrogram.

This procedure gives us K clusters of factors which now have to be transformed into K representative factors for each cluster. For the purposes of this paper, we average the underlying factors to give rise to a new representative factor for each cluster.

4.2.3. Factor Selection Through Shapley Values

Shapley values were originally introduced as "a value for n-person games" by Lloyd S. Shapley as a game-theoretical concept to determine the value of each contributor in a game in the presence of potential costs and benefits of alliances between contributors [37]. In 2017, Shapley values gained a surge of popularity in machine learning (ML) as a model-agnostic way of determining feature importances through the introduction of SHAP [33]. Without going into excessive detail, SHAP allows for the efficient calculation of feature importances based on Shapley values. These feature importances subsequently allow ML researchers to asses the importance of each feature in the presence of all other features.

Given a large universe of factors, we can now use SHAP to determine the importance of each factor in the presence of all other factors. Concretely, our SHAP based factor selection proceeds as follows:

1. Given a large universe of L factors, and a universe of N test-assets, construct the following linear factor model:

$$r^e_{i,t} = \alpha_i + \sum_{j=1}^{L} \beta_{j,i} f_{j,t} + \epsilon_{i,t} \tag{4.1}$$

[4] See Appendix C for the definition of this distance matrix.

Here, $r^e_{i,t}$ is the excess return of test-asset i at time t, $f_{j,t}$ is the return of factor j at time t, α_i and $\beta_{j,i}$ are the regression intercept and coefficients for test-asset i, and $\epsilon_{i,t}$ are the regression residuals.
2. For the above linear factor model, calculate the feature importance of each factor using SHAP and scale the resulting vector of feature importances such that they sum to 1.
3. Use the feature importance vector to select the relevant factors.

As with the other models, different selection criteria are possible. For the purposes of this paper, the K factors with the largest feature importance are selected.

4.3. Factor Engineering

This paper considers two distinct algorithms to construct new regime-dependent factors as outlined in Section 3.3. First, the Linear Boosting algorithm from Ilic, et al. [25] is used to build non-linear combinations and transformations of existing factors. Second, the RP-PCA algorithm proposed by Lettau, et al. [31] is used to find latent factors in the equity cross-section that are not captured by existing factors. By implementing RP-PCA and Linear Boosting, both linear and non-linear latent factor models are considered, respectively.

4.3.1. Linear Boosting

Consider the linear factor model from Equation 4.1. This model is used as the base learner in Linear Boosting as described in Section 3.3.1. This model is specified as a multi-output linear regression with the factors as independent variables and the asset cross-section as an N-dimensional dependent variable[5]. Following the Linear Boosting framework, the original factor set, $F \in \mathbb{R}^{T\times L}$, is extended with K new factors, $F' \in \mathbb{R}^{T\times K}$, resulting in a new factor set, $F \cup F' \in \mathbb{R}^{T\times(L+K)}$. This paper uses the `linear-tree` package from Cerliani [5] to implement Linear Boosting.

4.3.2. RP-PCA

RP-PCA, outlined in Section 3.3.2 is a special case as it can both act as a factor selection and a factor engineering tool. First, when applied over a cross-section of factors to extract the top K RP-PCA factors, it acts as a factor selection tool. This way a large cross-section of factors is reduced into K new factors which are linear combinations of the original factors and explain most of the risk and return of the original factors.

Next, RP-PCA can be applied over a cross-section of asset returns to extract the top K latent factors. In this case, the original factor set, $F \in \mathbb{R}^{T\times L}$, is extended with the K new factors, $F' \in \mathbb{R}^{T\times K}$, resulting in a new factors set, $F \cup F' \in \mathbb{R}^{T\times(L+K)}$. As there are potential similarities between these RP-PCA factors and the existing factors, it is advisable to perform a factor selection step afterwards to filter out similar factors. This paper implements RP-PCA by translating the Matlab source code from Lettau, et al. [31] into Python.

[5] Although the source paper uses LASSO as a base learner, the proposed factor model already incorporates factor selection procedures. As a result, this base learner is fitted through OLS to avoid a double factor selection.

4.4. Factor Blending

We build on top of the model proposed by Costa, et al. [9], by relaxing some assumptions and extending the dynamics of the model. (See Section 3.4 for an overview of the original model.) First, we do not require the factor returns to be mean-zero as we are interested in the potential returns-forecasting power of the factors, i.e. we let $f_{t,i} \sim N(\mu_{f_i}, F_i)$ where μ_{f_i} are the expected factor returns in state i. Next, we allow for different factor sets for each regime, i.e. $f_{t,i} \in \mathbb{R}^{L_i}$ where the number of factors in state i, L_i, and the set of factors is determined by the factor selection and engineering procedures outlined above. While we retain the factor model's specification from Eq. (3.2), the resulting output undergoes a slight modification. Let the asset expected excess return vector within a fixed state i be given by:

$$\mu_i = \alpha_i + K_i^T \mu_{f_i} \tag{4.2}$$

Then for a two-state model, the asset expected excess return vector and VCV conditional on being in state i today are given by:

$$\hat{\mu}_i^e = \gamma_{i,1}\mu_1 + \gamma_{i,2}\mu_2 \tag{4.3}$$

$$\begin{aligned}\hat{\Sigma}_i = {} & \gamma_{i,1}(K_1^T F_1 K_1 + D_1) + \gamma_{i,2}(K_2^T F_2 K_2 + D_2) \\ & + \gamma_{i,1}(1-\gamma_{i,1})\mu_1\mu_1^T + \gamma_{i,2}(1-\gamma_{i,2})\mu_2\mu_2^T \\ & - \gamma_{i,1}\gamma_{i,2}(\mu_1\mu_2^T + \mu_2\mu_1^T)\end{aligned} \tag{4.4}$$

Similar to the model from Costa, et al. [9] this model can easily be extended for more than two states, as shown in Appendix D.

4.5. Proposed Factor Model Overview

To summarize, the proposed factor model framework consists of four key components, each of which is optional, can be combined with the others, and has multiple implementations:

1. Regime Modeling
   (a) Gaussian Mixture Model HMMs over the Fama French market factor, Mkt-RF, are used to fit historical regimes and partition the data accordingly.
2. Factor Selection
   (a) Sequential factor selection based on the GRS statistic.
   (b) Factor selection based on Shapley values.
   (c) Factor selection based on agglomerative hierarchical clustering defined over factor covariances.
   (d) RP-PCA applied over factor time-series.
3. Factor Engineering
   (a) Linear Boosting applied over a linear factor model.
   (b) RP-PCA applied over asset return time-series.
4. Factor Blending

(a) A closed form model defined over regime-dependent factor models and regime transition-probabilities is used to provide a single asset expected return vector and VCV as the model output.

Evidently, factor blending hinges on the application of regime modeling. If factor selection and/or engineering are used in conjunction with regime modeling, then they are always applied independently to each regime.

This structure innovates on the existing body of literature in numerous ways. First, to our knowledge this is the first study where each of these components are considered jointly, using large factor models. This allows us to assess the power of each of these techniques in light of the availability of the others. Second, to our knowledge, Linear Boosting has not yet been applied to the problem of identifying novel, non-linear factors. Finally, by providing a highly structured model, the impact of each step can be isolated and explained ex-post, something that is highly relevant to industry applications.

§5. Data Description

Three datasets are employed within this paper to evaluate the proposed factor model from Section 4. First, the Fama-French market factor and risk free rate are retrieved from Kenneth French's Data Library [15]. The market factor is used to perform regime detection and the risk free rate is used to calculate excess returns. Next, the full CRSP dataset is retrieved from WRDS [44]. This dataset offers a large cross-section of daily US-based equity returns which will constitute the test assets during model evaluation and backtesting. Finally, the US large factor model (USFM) recently open-sourced by Jensen, et al. [26] is retrieved as the principal large factor model used within this paper.

In the remainder of this section, the key data cleaning steps performed for this research are outlined. The data retrieved from Kenneth French's Data Library did not show any inconsistencies or problems and will hence not be discussed in this section. Further discussion on universe selection are deferred to Section 6 when the empirical framework is explained.

5.1. Asset Returns (CRSP)

The CRSP dataset provides daily returns on nearly every asset listed on US stock exchanges from January 1990 until December 2023. The CRSP dataset offers multiple ways of identifying assets. For this report, each `permno` is considered a separate asset. As the `permno` of a stock tends to change after major corporate actions (e.g. spin-offs or mergers), this avoids having to deal with those complexities. Furthermore, the `curcdd` field is used to filter for USD-denominated assets only and the `loc` field is used to filter for US-headquartered firms only. Finally, the `linkprim` flag is used to filter out any non-primary shares and the `exchg` flag is used to filter out any OTC shares. These high-level filters ensure that only reasonably tradable, US-based assets are considered within this paper.

In addition to identifying and selecting the relevant assets, several pre-processing steps are required to ensure the reliable data for these assets. First, the CRSP dataset does not offer returns data directly, only price data. To simplify the analysis, this paper assumes

that all dividends are instantly re-invested into an asset. Therefore, we need to calculate the total returns adjusted for dividends and stock splits, which we will simply refer to as *adjusted returns* henceforth. To this end, CRSP provides three important fields:

- `prccd`: daily closing price of the asset.
- `trfd`: daily adjustment factor of the asset.
- `ajexdi`: daily ex-dividend adjustment of the asset.

Using these three field, adjusted returns can be calculated as follows. Let:

$$p_t^a = \texttt{trfd}_t^a \cdot \texttt{prccd}_t^a / \texttt{ajexdi}_t^a$$

then the returns for asset a at time t are given by:

$$r_t^a = \frac{p_t^a - p_{t-1}^a}{p_{t-1}^a}$$

Missing values for `trfd` and `ajexdi` are handled in two ways. First, missing values at the very beginning of each asset time-series are filled with 1. Underlying is the assumption that missing data at the start of the time series simply represents the absence of any adjustments. Second, remaining missing values are forward filled. Underlying this is the assumption that these adjustment factors are highly persistent so in the absence of any data, the previous available value is the next best guess. Missing values in `prccd` are not handled as there is no reasonable way of knowing the true value[6].

Next to returns, reliable market caps are necessary for subsequent universe selection. To this end, CRSP provides the field `cshoc`: the shares outstanding at the close of every trading day. Market caps can then simply be calculated at the close of each trading day t for each asset a as:

$$\texttt{mcap}_t^a = \texttt{cshoc}_t^a \cdot \texttt{prccd}_t^a$$

Missing values in `cshoc` are handled by forward filling the data. Underlying is the assumption that shares outstanding is highly persistent so in the absence of any data, the previous available value is the next best guess. Missing values at the start of each asset time-series are not handled as there is no way of reasonably knowing the true value and backward filling could constitute a potential look-ahead bias.

As seen on Figure 2, after undertaking these pre-processing steps only a small fraction of the returns data is missing each month. Market cap data, however, is quite sparse prior to 1997 as seen on Figure 3. This means that prior to 1997, no reliable universe selection can be performed based on market cap. As it turns out, this issue is negligible as the proposed factor model requires a certain burn-in period and hence universe selection only starts as of 2005.

[6] Forward filling this data would lead to a substantial measurement error while backward filling would constitute a look-ahead bias.

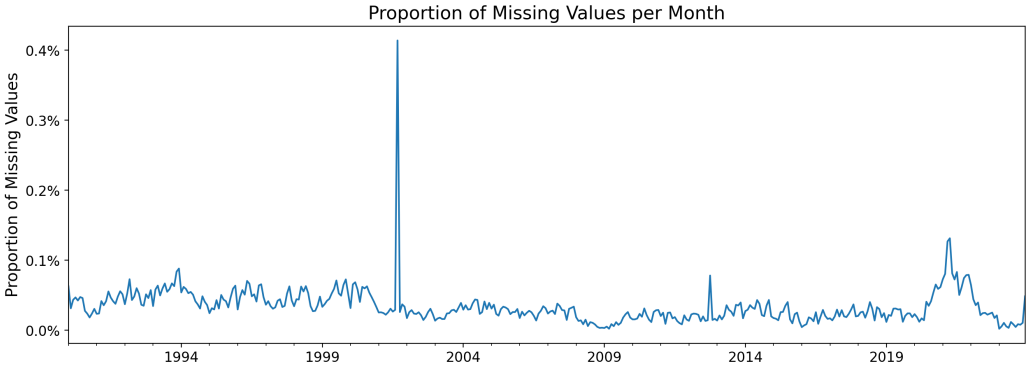

Figure 2. : Proportion of missing returns data over time.

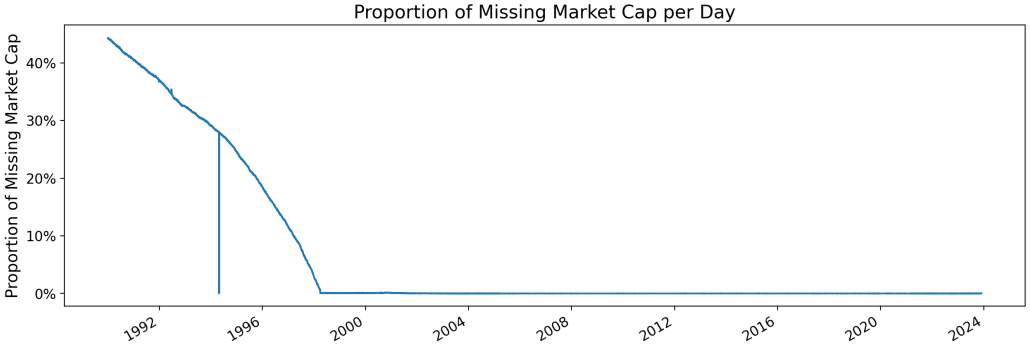

Figure 3. : Proportion of missing market cap data over time.

When inspecting the resulting asset returns, at first sight it appears that there are outliers in the extreme positive and negative returns. Nonetheless, a qualitative analysis clarified that the bulk of such outliers were in fact real events, often related to periods of low liquidity, micro-caps, and market crashes. As a result, we opt not to winsorize or clip the returns for backtesting purposes. Under this approach, we have to accept that a handful of the asset returns were true outliers, but on the flip-side we do not mechanically under- or overstate results by removing extreme, but real asset returns.

A final consideration for the asset returns is how to deal with new arrivals, e.g. IPOs, and market exits, e.g. bankruptcies, delistings, and acquisitions. From a pre-processing standpoint, no action is taken here as these are all considerations for the backtesting framework and model fitting. The model fitting takes a simplistic, but tradable approach to this problem, namely by only considering assets that have full-history data available over the model's training window. Although strict, this approach avoids that the results are clouded by biases introduced by handling missing data. On the flip side, it constitutes a certain

selection bias, namely only assets with full-history are considered. Nonetheless, this bias is not forward looking and applies equally to each model, making the results comparable[7].

5.2. Global Factor Model (GFM) and US Factor Model (USFM)

The open-source factor models from Jensen, et al [26] are already of impeccable quality with no apparent missing or false data. Nonetheless, it is important to understand this dataset as it is crucial to the results of this paper. To that end we compare the provided Global Factor Model (GFM) with the US Factor Model (USFM).

The GFM contains data on 153 factors across 93 countries [26]. For each factor, Jensen, et al. construct the 1-month holding period return for each country. The factor returns are defined through a high minus low tercile sort on the underlying signal, corresponding to the excess return of a long-short, zero-net-investment strategy. Each factor is long (short) the tercile identified by the original paper to have the highest (lowest) expected return. Multiple weighting schemes are provided. This paper uses capped value weighted returns, following the recommendation from Jensen, et al. [26].

The GFM and the USFM are identical until 1983. This is purely mechanical as new countries only enter the dataset post this date. Next to the number of countries, the number of factors varies over time. This is also largely mechanical based on when the necessary data becomes available. Most importantly, as of 1990 high-quality data is available for all 153 factors in the USFM, in line with the available data on the asset returns.

Important to note is that the factor models report daily excess returns and not raw returns. As a result, it is not evident to simply calculate compounded returns from these time series. Jensen, et al. [26] also provide monthly factor returns, but frequencies other than daily or monthly have to be reconstructed from the bottom up using the daily portfolio returns.

Anecdotally, when considering the logarithmic sums of the excess return of both models over time, we observe that the USFM outperforms the Global Factor Model in about 60% of the factors. More generally, Swade, et al. [39] show that the USFM is most appropriate when considering US stock universes whereas for other countries the GFM performs better than the country-specific factor model. We follow this suggestion and only consider the USFM for the remainder of this paper.

§6. Empirical Framework

Given the model pipeline and necessary data, the performance of the model can be evaluated. Conceptually, this happens from two angles. First, the actual model outputs are evaluated, i.e. how well does the estimated asset expected excess return vector and VCV correspond to the realized asset expected excess return vector and VCV over a future period? Although this exercise allows for a pure model evaluation, a core problem is that the true expected excess return vector and VCV are latent variables. By using the

[7] In a production model, different approaches could be taken to deal with new arrivals, e.g. by using industry returns or similar as a proxy to fill out missing historical data.

empirical realizations as a proxy for the true target, this analysis is implicitly biased.

In light of these concerns, the factor model pipeline is also evaluated by simulating a real-world application, namely portfolio optimization. Through this exercise, the performance of each factor model configuration is evaluated based on the resulting portfolio characteristics such as turnover and various risk-return metrics. Figure 4 provides a visual overview of where in the portfolio optimization process each evaluation step takes place.

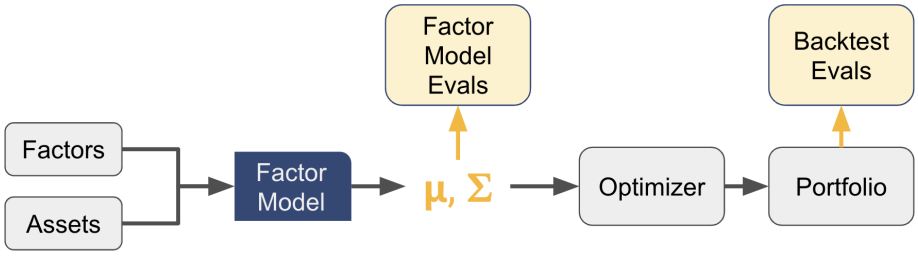

Figure 4. : Diagram outlining where in the portfolio optimization pipeline each model evaluation step takes place.

The remainder of this section will outline how both the pure model evaluation and portfolio backtest are setup. Section 7 will subsequently present and draw conclusions on the results of both exercises.

6.1. Pure Model Evaluation

The pure model evaluation is designed to test the model's ability to forecast future returns and observed covariances of assets, under different model configurations. To this end, each model is fitted and evaluated over rolling windows as shown in Figure 5. Different evaluation metrics can be considered on the expected return vector and VCV, both individually and jointly.

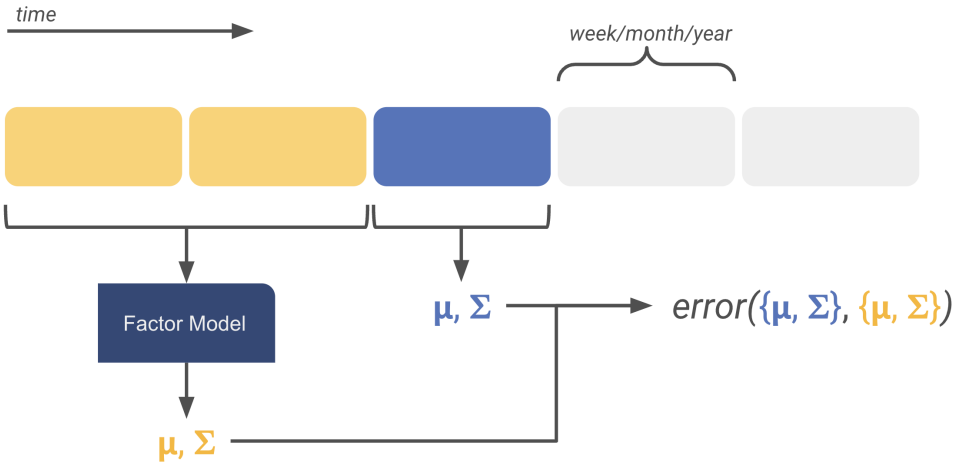

Figure 5. : Factor model evaluation diagram

The output of this evaluation framework are time-series of evaluation metrics for different model configurations. These can subsequently be analyzed to assess both the average as well as the time-varying performance of each model. It is important to realize that the both are equally important. Depending on the use case, a model with worse average performance, but great performance during specific important periods may be preferred over a model with the opposite characteristics.

Different evaluation metrics are considered between the expected and realized return vector and VCV, both individually and jointly. First, Mean Squared Error (MSE) judges the size of potential errors variance-covariance matrices (i.e. the Frobenius Norm) and the expected return vectors (i.e. the 2-Norm). Second, since many optimization problems are invariant under scale and bias of the expected returns, Pearson and Kendall correlation between the expected and realized return vectors measure the correct direction or ranking of the forecasts. We also consider the MSE and correlations of the VCV diagonals to isolate potential performance differences between variances and covariances. Finally, 2-Wasserstein distance is used as a joint metric over expected returns and the VCV[8]. This metric measures the similarity between the mutltivariate normal distribution of asset returns implied by the factor model and the one implied by realized returns.

Looking towards the proposed model overview of Section 4.5, there are much more possible combinations of models than can be reasonably be evaluated within this paper, especially when considering all hyperparameters and training period sizes. As a result, a representative set of model configurations is selected, with results presented in Section 7. To evaluate different model configurations, we benchmark the performance against a static factor model and a regime-dependent model without any factor selection nor engineering.

6.2. Portfolio Optimization

Next to the pure model evaluation, we also want to understand the potential benefits of our proposed model in real-world applications. To that end, we apply standard portfolio optimization on our factor models to evaluate the performance of the resulting portfolios. Concretely, we consider Mean-Variance Optimization (MVO) and Minimum Volatility. To evaluate the resulting portfolios, we apply these two optimization techniques over two benchmark factor models. First, we consider the static US Factor Model (USFM)[9] from Jensen, et al. [26]. In line with Swade, et al. [39], we enhance the USFM with the Fama-French market factor, Mkt-RF [15]. From hereon, we simply refer to this enhanced model as the USFM. Second, we consider a dynamic version of the USFM where we apply 2-state regime conditioning, but no feature selection or engineering. Through these two benchmarks, we aim to isolate the source of potential improvements to one or a combination of regime conditioning, factor selection, and factor engineering. Next to these two optimized portfolios, a simple equal-weighted, monthly rebalanced portfolio is also included as a model-agnostic benchmark.

[8] Under a normality assumption, He [22] shows that there is a closed form solution to calculate this distance metric.

[9] As our equity universe consists of US stocks, the US Factor Model explains the equity cross-section better than the Global Factor Model as shown by Swade, et al. [39].

To go from portfolio optimization to tradable portfolios, we consider portfolios rebalanced at the beginning of each business month. For each configuration of factor model and portfolio optimizer we repeat the following on each rebalance date:

1. Select the 200 stocks with the largest market cap at the close of the previous month's last trading day.
2. Fit the chosen factor model on a 15-year lookback window[10,11] for the selected assets.
3. Extract the relevant expected return vector (μ) and asset VCV (Σ) from the factor model.
4. Calculate the optimal weights using μ and Σ with the chosen portfolio optimizer under no-short selling, zero leverage, and full investment constraints.
5. Simulate the returns of this portfolio over the month.

This optimization exercise is applied over the top 200 stocks by market cap for three reasons. First, the data quality on large caps is much better than on small caps, mitigating the impact of outliers and other data issues. Second, it is infeasible to perform quadratic optimization over very large asset universes (e.g. 1000+ stocks), because of computational constraints and increased overfitting. Finally, long-only constrained portfolios have the undesirable tendency to create non-diversified portfolios with many (near) zero asset weights [7][18][32]. In the absence of constraints or other methods to deal with these issues, there is little benefit from considering a large asset universe for the optimization problem. Regardless of this choice, there is nothing inherent to our proposed model that disallows anyone from applying it over different asset universes.

In general, these portfolios could be subject to any constraints that a typical quadratic optimizer allows for. For the purposes of this paper, we apply three constraints on every portfolio: no short-selling, zero-leverage, and full investment. These constraints form the basis of, but are weaker than what most portfolio managers face in practice. As Clarke, et al. [6] argue, adding further constraints to the optimization problem reduces the ex-post information ratio. Nonetheless, using a better factor model will still benefit portfolio managers, even under constraints as argued by Ledoit, et al. [30].

After running the backtests for different model configurations, different performance metrics are reported, for three key purposes:

1. **Risk-Return Metrics**: Are these portfolios attractive for prospective investors? Do they have superior tail-risk characteristics?
2. **Diversification Metrics**: How concentrated are these portfolios?
3. **Turnover Metrics**: How costly is it to implement these portfolios?

Similar to the pure model evaluation, there are much more possible combinations of models than can be reasonably be evaluated within this paper. As a result, a representative set of model configurations is selected, with results presented in Section 7.

[10] The choice of a 15 year lookback window is in line with our previous industry project [11] and ensures that the regime model is exposed to multiple different market scenarios.

[11] As mentioned earlier, the Fama-French [15] market factor, Mkt-RF, is used for regime modeling and the USFM [26] is used as the factor model input.

§7. Results

7.1. Model Evaluation

To avoid an information overload, this section will focus on the key insights[12] gained from evaluating many model configurations through the metrics defined in Section 6.1. Evaluation tables backing up these insights can be found in Appendix E. All metrics were computed across different validation window sizes using a training set of 3,750 business days with a 62-day stride. Furthermore, all metrics were computed over the portfolio optimization universe described in Section 6.2.

A first insight is that regime-dependent factor models predict the realized return distribution better in the short term (1-3 months) than in the long term (1-3 years). This is shown clearly through the respective Wasserstein distances and holds both for 2- and 3-state GMMHMMs. Notably, under a 2-state model, neither factor selection, nor factor engineering has a strong impact on the Wasserstein distance. However, under a 3-state model, these steps have a bigger impact. On the selection side, sequential selection through GRS deteriorates performance, whereas Shapley based selection improves performance. On the engineering side, RP-PCA deteriorates results, whereas Linear Boosting improves results, especially over the long term (1-3 years). Nonetheless, the standard deviation over these metrics prevent us from drawing strong conclusions, especially for the regime-dependent models.

The large standard errors in the regime-dependent model metrics lead us to our next insight. As seen in Figure 6, the performance of static factor models is much more stable than that of regime-dependent models. Even more interesting is that the regime-dependent models seem to do especially well during "stable" times, e.g. 2012-2016 (as also noted during our previous industry project [11]). A main driver behind this phenomenon is that regime-dependent models can properly condition themselves on the stable state and hence are not overly influenced by historically challenging conditions. On a static factor model, however, historically challenging conditions (and outliers) can significantly impact the estimated expected factor returns and covariances.

The improved performance over the stable state is not a free lunch however, as can be seen through the spikes of erroneous behavior in Figure 6. There could be multiple reasons behind this issue. First, there is the obvious problem of misclassification. These regime models are not perfect and as such, misclassifying today's state will lead to bad estimates for the asset expected return vector and VCV. Second, imagine that the regime model splits data into a stable and abnormal state. Since the abnormal state occurs less frequency, its parameters may be less efficiently estimated, leading to a bad forecast for this state. Third, and most subtle, is the choice of factor blending (see Section 4.4). This model is fundamentally a 1-period model, i.e. the blended factor model is optimal for tomorrow given today's regime information. However, we are now evaluating these factor models over longer periods, i.e. 1-3 months. This highlights a potential improvement on the proposed factor model.

[12] Metrics which did not show novel insights or variation between models are not reported but can be found in the online appendix to this paper [12].

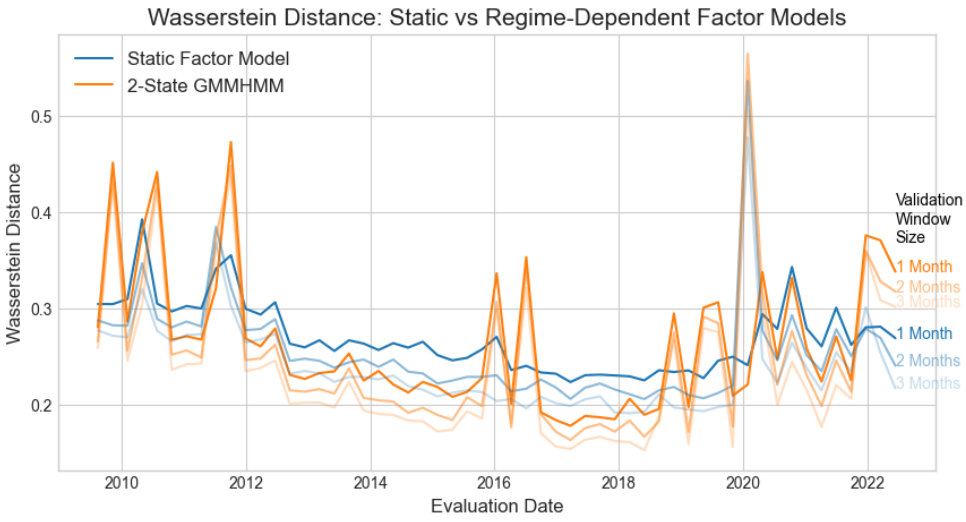

Figure 6. : Evolution of 2-Wasserstein distance over time over different validation periods. The orange lines are the performance for a regime-dependent model with 2-state GMMHMM and no factor selection or engineering. The blue lines are the performance of a static USFM benchmark model.

Another critical takeaway is the superiority of regime-dependent factor models to directional returns forecasting. To see this, we have to look at the Pearson correlation between expected and realized returns. First, on a 1-month frequency, the static factor model has a correlation of -1.4% on average between expected and realized returns, whereas a 2-state GMMHMM-based model has a correlation of +5.1% and a 3-state GMMHMM-based model has a correlation of +1.9% on average. Second, 2-state models outperform static models on all evaluation windows and 3-state models outperform on all but one evaluation windows. Next, similar to the Wasserstein distance comparison, factor selection and engineering have little to no impact on 2-state models. However, they have a strong impact on 3-state models. On the factor selection side, both Shapley and RP-PCA show significant improvements on 3-state models. On the factor engineering side, RP-PCA deteriorates results, whereas Linear Boosting shows some improvements for short-term forecasts (1 month to 1 year).

This analysis emphasizes the value of regime-dependent models in conditions where market regimes are distinctly identifiable, especially for forecasting return-dynamics more accurately during stable market conditions. Nonetheless, it is clear that neither dynamic nor static models offer a one-size-fits-all solution. Both the investment horizon and investors' preference for stable models influence which model is the most appropriate. The number of states in a regime model also play an important role in finding the appropriate model. Based on this analysis, 3-state models appear to be more sensitive to factor selection and engineering than 2-state models.

7.2. Portfolio Backtests

Factor Model	Regime Model	Factor Selection	Factor Engineering	Select First?
USFM	/	/	/	/
USFM	GMMHMM (2)	/	/	/
USFM	GMMHMM (2)	Hierarchical	/	/
USFM	GMMHMM (2)	Shapley	/	/
USFM	GMMHMM (2)	GRS	/	/
USFM	GMMHMM (2)	RP-PCA	/	/
USFM	GMMHMM (2)	/	RP-PCA	/
USFM	GMMHMM (2)	/	EBLR	/
USFM	GMMHMM (2)	Hierarchical	RP-PCA	Yes
USFM	GMMHMM (2)	Shapley	RP-PCA	Yes
USFM	GMMHMM (2)	Hierarchical	EBLR	Yes
USFM	GMMHMM (2)	Shapley	EBLR	Yes
USFM	GMMHMM (2)	Hierarchical	RP-PCA	No
USFM	GMMHMM (2)	Shapley	RP-PCA	No
USFM	GMMHMM (2)	Hierarchical	EBLR	No
USFM	GMMHMM (2)	Shapley	EBLR	No

Table 1.: Overview of all models considered within the portfolio backtests. Each model uses the USFM as its underlying large factor model and a 2-state Gaussian Mixture Model HMM over Mkt-RF as the regime model. Each factor selection procedure selects 15 factors and each factor engineering procedure constructs 10 factors, unless the engineering step occurs first. In that case, 10 factors are constructed, but 25 are selected. This ensures that all models that perform both selection and engineering consider the same amount of final factors, regardless of the order in which these steps took place. Also note that the first two rows represent the static and regime-dependent benchmark factor models, respectively.

The portfolio backtests prescribed in Section 6.2 are ran from 2005-2023 for each model in Table 1. In line with the results from Section 7.1, 2-regime models are considered as their performance aligns the most with a 1 month investment objective. An equal weighted portfolio is also included as a model-agnostic benchmark. The wide variety of model configurations under consideration enable us to identify exactly what components are beneficial or redundant to construct optimal portfolios.

7.2.1. Portfolio Performance

When looking at the portfolio performance, it is important to consider the objectives of each component of the proposed factor model. On the one hand, regime-conditioning and factor engineering aim at improving the overall performance of optimized portfolios. This can be realized either in better tail-risk measures or simply overall risk-return characteristics. On the other hand, factor selection aims at maintaining the performance of large factor models, while avoiding the use of hundreds of factors.

Given this context, Tables 2 and 3 show a broad set of performance metrics for the optimized portfolios. First note that the equal weighted (EW) and the two static benchmarks are nearly identical in terms of Sharpe and Sortino ratio. Next to this, the static benchmarks are more left-skewed and more leptokurtic than the EW benchmark. Although the static benchmarks do show better drawdown measures, they are not volatility matched to the EW benchmark. Under volatility matching, they fail to show much better drawdown measures. Peeking ahead to Section 7.2.2, the static portfolios also show a higher TO and less diversification than the EW benchmark.

Based on these observations, the static large factor model is not very attractive relative to the simple EW portfolio. The question now becomes whether regime-dependency can fix these issues.

Comparing mean-variance optimization to minimum volatility portfolios, MVO appears to be the better choice overall, while minimum volatility offers slightly better tail-risk protection as evidenced by the MDD and mean annual MDD. Comparing regime-dependent models to their static counterpart, regime-dependent models are superior across the board. They lead to higher returns, less volatility, less tail-risk, and an overall better risk-return trade off. Remarkably, the regime-dependent portfolios have a less negative skewness and smaller Fisher kurtosis, indicating that their returns were more normal over this backtest than those of their static counterparts. Furthermore, this outperformance also largely holds relative to the EW benchmark, with the exception of a slightly more negative skewness.

Looking towards factor selection, it appears to deliver on its objective. There are only negligible differences in the portfolio performance between the regime-dependent models with and without factor selection. This has two important implications. First, this suggests that many of the factors in the USFM are redundant, at least in one of the regimes. Second, including redundant factors is not necessarily detrimental to the portfolio performance as factor selection does not lead to improved results. This opens up interesting modeling choices in practice. On the one hand, small factor models are easier to understand and have less numerical complexity. On the other hand, incorporating redundant factors to satisfy client demands may not lead to worse portfolios overall[13].

[13] For instance, clients may have certain exposure constraints that can only be satisfied by incorporating redundant factors or they may simply wish to see the optimal portfolio expressed in terms of factors they are familiar with (rather than the just the optimal factor set).

Strategy	Regime Dependent	Selection	Engineering	Select First?	Mean Return	Vol	Sharpe Ratio	Sortino Ratio	MDD
EW*	/	/	/	/	0.97%	4.41%	0.20	0.27	-46.52%
MVO*	No	/	/	/	0.80%	3.39%	0.20	0.28	-33.90%
MVO	Yes	/	/	/	0.90%	3.24%	0.24	0.35	-32.25%
MVO	Yes	Hierarchical	/	/	0.90%	3.26%	0.24	0.35	-32.44%
MVO	Yes	Shapley	/	/	0.90%	3.26%	0.24	0.35	-32.25%
MVO	Yes	GRS	/	/	0.90%	3.27%	0.24	0.35	-32.00%
MVO	Yes	RP-PCA	/	/	0.91%	3.25%	0.24	0.36	-32.19%
MVO	Yes	/	RP-PCA	/	0.89%	3.25%	0.24	0.35	-32.22%
MVO	Yes	/	EBLR	/	0.90%	3.26%	0.24	0.35	-32.44%
MVO	Yes	Hierarchical	RP-PCA	Yes	0.89%	3.27%	0.24	0.34	-32.94%
MVO	Yes	Hierarchical	RP-PCA	No	0.90%	3.25%	0.24	0.36	-32.46%
MVO	Yes	Shapley	RP-PCA	Yes	0.90%	3.25%	0.24	0.36	-32.44%
MVO	Yes	Shapley	RP-PCA	Yes	0.90%	3.26%	0.24	0.35	-32.11%
MVO	Yes	Hierarchical	EBLR	Yes	0.90%	3.26%	0.24	0.35	-32.26%
MVO	Yes	Hierarchical	EBLR	No	0.90%	3.26%	0.24	0.35	-32.40%
MVO	Yes	Shapley	EBLR	Yes	0.90%	3.25%	0.24	0.36	-32.40%
MVO	Yes	Shapley	EBLR	No	0.90%	3.25%	0.24	0.35	-32.40%
Min Vol*	No	/	/	/	0.78%	3.43%	0.20	0.27	-31.95%
Min Vol	Yes	/	/	/	0.87%	3.20%	0.24	0.34	-29.10%
Min Vol	Yes	Hierarchical	/	/	0.87%	3.20%	0.24	0.34	-29.31%
Min Vol	Yes	Shapley	/	/	0.87%	3.20%	0.24	0.34	-29.10%
Min Vol	Yes	GRS	/	/	0.87%	3.20%	0.24	0.34	-29.10%
Min Vol	Yes	RP-PCA	/	/	0.87%	3.20%	0.24	0.34	-29.11%
Min Vol	Yes	/	RP-PCA	/	0.87%	3.20%	0.24	0.34	-29.15%
Min Vol	Yes	/	EBLR	/	0.87%	3.20%	0.24	0.34	-29.31%
Min Vol	Yes	Hierarchical	RP-PCA	Yes	0.86%	3.21%	0.23	0.34	-29.67%
Min Vol	Yes	Hierarchical	RP-PCA	No	0.88%	3.20%	0.24	0.34	-29.21%
Min Vol	Yes	Shapley	RP-PCA	Yes	0.87%	3.20%	0.24	0.34	-29.31%
Min Vol	Yes	Shapley	RP-PCA	Yes	0.87%	3.21%	0.24	0.34	-29.14%
Min Vol	Yes	Hierarchical	EBLR	Yes	0.87%	3.20%	0.24	0.34	-29.11%
Min Vol	Yes	Hierarchical	EBLR	No	0.88%	3.20%	0.24	0.34	-29.27%
Min Vol	Yes	Shapley	EBLR	Yes	0.87%	3.20%	0.24	0.34	-29.27%
Min Vol	Yes	Shapley	EBLR	No	0.88%	3.20%	0.24	0.34	-29.27%

Table 2.: Core summary statistics on the optimized portfolios' performance. The benchmark models are indicated with a star (*), while the remaining rows represent the other models from Table 1. All reported numbers are monthly, except for MDD which is full-sample.

Strategy	Regime Dependent	Selection	Engineering	Select First?	Skew	Fisher Kurtosis	Max Monthly Loss	Max Monthly Gain	Mean Annual MDD
EW*	/	/	/	/	-0.42	1.50	-16.26%	14.06%	-12.28%
MVO*	No	/	/	/	-0.65	1.92	-13.57%	10.29%	-8.99%
MVO	Yes	/	/	/	-0.54	1.32	-11.61%	9.56%	-8.42%
MVO	Yes	Hierarchical	/	/	-0.56	1.36	-11.61%	9.56%	-8.48%
MVO	Yes	Shapley	/	/	-0.56	1.35	-11.61%	9.56%	-8.45%
MVO	Yes	GRS	/	/	-0.56	1.35	-11.61%	9.75%	-8.43%
MVO	Yes	RP-PCA	/	/	-0.53	1.29	-11.61%	9.75%	-8.40%
MVO	Yes	/	RP-PCA	/	-0.56	1.35	-11.61%	9.56%	-8.46%
MVO	Yes	/	EBLR	/	-0.56	1.35	-11.61%	9.56%	-8.46%
MVO	Yes	Hierarchical	RP-PCA	Yes	-0.59	1.49	-12.06%	9.56%	-8.59%
MVO	Yes	Hierarchical	RP-PCA	No	-0.54	1.29	-11.61%	9.56%	-8.45%
MVO	Yes	Shapley	RP-PCA	Yes	-0.54	1.31	-11.61%	9.56%	-8.43%
MVO	Yes	Shapley	RP-PCA	Yes	-0.56	1.36	-11.61%	9.56%	-8.43%
MVO	Yes	Hierarchical	EBLR	Yes	-0.56	1.36	-11.61%	9.56%	-8.46%
MVO	Yes	Hierarchical	EBLR	No	-0.57	1.36	-11.61%	9.56%	-8.48%
MVO	Yes	Shapley	EBLR	Yes	-0.54	1.29	-11.61%	9.56%	-8.44%
MVO	Yes	Shapley	EBLR	No	-0.54	1.35	-11.61%	9.75%	-8.42%
Min Vol*	No	/	/	/	-0.71	1.99	-13.60%	9.37%	-9.19%
Min Vol	Yes	/	/	/	-0.57	1.33	-11.81%	8.66%	-8.12%
Min Vol	Yes	Hierarchical	/	/	-0.57	1.33	-11.81%	8.66%	-8.10%
Min Vol	Yes	Shapley	/	/	-0.57	1.33	-11.81%	8.66%	-8.06%
Min Vol	Yes	GRS	/	/	-0.57	1.34	-11.81%	8.70%	-8.06%
Min Vol	Yes	RP-PCA	/	/	-0.57	1.34	-11.81%	8.70%	-8.07%
Min Vol	Yes	/	RP-PCA	/	-0.57	1.34	-11.81%	8.66%	-8.07%
Min Vol	Yes	/	EBLR	/	-0.57	1.33	-11.81%	8.66%	-8.09%
Min Vol	Yes	Hierarchical	RP-PCA	Yes	-0.59	1.45	-12.31%	8.66%	-8.15%
Min Vol	Yes	Hierarchical	RP-PCA	No	-0.58	1.34	-11.81%	8.66%	-8.08%
Min Vol	Yes	Shapley	RP-PCA	Yes	-0.57	1.34	-11.81%	8.66%	-8.08%
Min Vol	Yes	Shapley	RP-PCA	Yes	-0.59	1.37	-11.81%	8.66%	-8.15%
Min Vol	Yes	Hierarchical	EBLR	Yes	-0.58	1.34	-11.81%	8.66%	-8.07%
Min Vol	Yes	Hierarchical	EBLR	No	-0.57	1.34	-11.81%	8.66%	-8.08%
Min Vol	Yes	Shapley	EBLR	Yes	-0.58	1.34	-11.81%	8.66%	-8.10%
Min Vol	Yes	Shapley	EBLR	No	-0.57	1.35	-11.81%	8.70%	-8.06%

Table 3.: Additional summary statistics on the optimized portfolios' performance. The benchmark models are indicated with a star (*), while the remaining rows represent the other models from Table 1. All reported numbers are monthly, except for mean annual MDD. The mean annual MDD captures the average MDD that an investor would observe over any 12 month period of holding the portfolio.

Finally, unlike factor selection, factor engineering fails to achieve its objective as there are only negligible differences in the portfolio performance between the regime-dependent models with and without factor engineering. There could be multiple reasons behind this failure to improve results. First, the USFM presents an ambitious benchmark to beat as its 153 factors are the result of decades of empirical finance research. As such, it is conceptually understandable that both RP-PCA and EBLR fail to uncover latent factors that were not captured yet. Second, the universe over which this backtest applies, the top 200 US stocks by market cap, is highly liquid and considered highly efficient. This further sets a high

bar as many factors are often attributed to low-liquidity and small cap stocks. Finally, the successful integration of factor engineering may require a different modeling approach. In the current model set-up, idiosyncratic returns (or "alpha") and idiosyncratic volatility are simply accounted for by a linear intercept and the in-sample residuals, as seen in Equations 4.2 and 4.4. Through this standard model, the idiosyncratic components may outshine any improvement on the systematic components.

7.2.2. Portfolio Weights

Beyond from understanding the performance of each portfolio, it is important to understand the asset weights behind these portfolios. Each portfolio is rebalanced at the beginning of the month and considers the same universe of assets. As seen in Table 4, the equal weighted portfolio invests in 154 assets each month, on average[14]. The optimized portfolios however are much more concentrated. As mentioned before, long-only constrained portfolios have the undesirable tendency to create non-diversified portfolios with many (near) zero asset weights [7][18][32]. MVO portfolios appear better diversified than minimum volatility portfolios. More importantly, portfolios from regime-dependent models are more diversified than their static counterparts.

The better diversification of regime-dependent models is not a free lunch, however. Note that the regime-dependent models cause a significantly higher turnover than their static counterparts. As seen by the equal weighted portfolio, this higher turnover cannot simply be attributed to holding a larger number of assets on average. Consider instead Equations 4.3 and 4.4, which show how the asset expected excess return vector and VCV are conditional on today's state. As a result, the optimal portfolio weights will also be conditional on today's state and state changes are likely to induce a larger turnover.

Finally, as far as the factor selection and engineering procedures go, there are little to no changes to the portfolio weights under any of the different configurations, in line with observed portfolio performance. This implies two important observations. On the one hand, the factor selection appears to do exactly what it is meant to do. It is able to significantly reduce the amount of factors considered in each state from 153 to 15 with little to no loss in performance. On the other hand, the factor engineering fails to make an impact. Neither the linear factors found by RP-PCA nor the non-linear factor found by linear boosting improve or even change the portfolio weights. Hence the portfolio weights do not justify incorporating any of these new factors into the model.

Next to diversification and turnover, it is also important to investigate the similarity of the optimal portfolio weights. As shown on Figure 7, The portfolio weights correlate with each other in four distinct clusters: static MVO, static minimum volatility, regime-dependent MVO, and regime-dependent minimum volatility. Note also how the two static models are relatively highly correlated between each other, but less correlated to the regime-dependent models. In line with the other observations on the portfolio weights, factor

[14] Although the universe considers the top 200 stocks by market cap, some assets are not selected since they did not have complete data over the past 15 years. This additional selection step aligns the equal weighted universe with the optimized portfolios which operate under this full-history requirement.

selection achieves its objective both under MVO and minimum volatility, with near perfect correlation to the regime-dependent model without factor selection. Similarly, factor engineering fails to achieve its objective with seemingly no impact on the portfolio weights.

Strategy	Regime Dependent	Selection	Engineering	Select First?	Mean # Assets	Mean Weight Change (%)	Monthly Turnover
EW*	/	/	/	/	154.2	0.21%	0.39
MVO*	/	/	/	/	38.7	0.26%	0.49
MVO	Yes	/	/	/	44.4	0.37%	0.68
MVO	Yes	Hierarchical	/	/	44.5	0.37%	0.68
MVO	Yes	Shapley	/	/	44.5	0.37%	0.68
MVO	Yes	GRS	/	/	44.4	0.37%	0.68
MVO	Yes	RP-PCA	/	/	44.5	0.37%	0.68
MVO	Yes	/	RP-PCA	/	44.4	0.37%	0.68
MVO	Yes	/	EBLR	/	44.5	0.37%	0.68
MVO	Yes	Hierarchical	RP-PCA	Yes	44.4	0.37%	0.69
MVO	Yes	Hierarchical	RP-PCA	No	44.4	0.37%	0.68
MVO	Yes	Shapley	RP-PCA	Yes	44.4	0.37%	0.68
MVO	Yes	Shapley	RP-PCA	No	44.3	0.37%	0.68
MVO	Yes	Hierarchical	EBLR	Yes	44.5	0.37%	0.68
MVO	Yes	Hierarchical	EBLR	No	44.5	0.37%	0.68
MVO	Yes	Shapley	EBLR	Yes	44.4	0.37%	0.68
MVO	Yes	Shapley	EBLR	No	44.4	0.37%	0.68
Min Vol*	/	/	/	/	26.7	0.37%	0.49
Min Vol	Yes	/	/	/	31.7	0.48%	0.66
Min Vol	Yes	Hierarchical	/	/	31.7	0.48%	0.65
Min Vol	Yes	Shapley	/	/	31.7	0.48%	0.65
Min Vol	Yes	GRS	/	/	31.6	0.48%	0.65
Min Vol	Yes	RP-PCA	/	/	31.7	0.48%	0.65
Min Vol	Yes	/	RP-PCA	/	31.7	0.47%	0.65
Min Vol	Yes	/	EBLR	/	31.7	0.48%	0.65
Min Vol	Yes	Hierarchical	RP-PCA	Yes	31.6	0.48%	0.66
Min Vol	Yes	Hierarchical	RP-PCA	No	31.7	0.48%	0.65
Min Vol	Yes	Shapley	RP-PCA	Yes	31.7	0.48%	0.65
Min Vol	Yes	Shapley	RP-PCA	No	31.6	0.48%	0.66
Min Vol	Yes	Hierarchical	EBLR	Yes	31.6	0.48%	0.65
Min Vol	Yes	Hierarchical	EBLR	No	31.7	0.48%	0.65
Min Vol	Yes	Shapley	EBLR	Yes	31.6	0.48%	0.65
Min Vol	Yes	Shapley	EBLR	No	31.6	0.48%	0.65

Table 4.: Summarizing statistics on the optimal portfolio weights. The benchmark models are indicated with a star (*), while the remaining rows represent the other models from Table 1. The mean number of assets indicates in how many assets each portfolio invests on average, and acts as a diversification metric. The mean weight change indicates the average weight change for assets with non-zero weights before or after the rebalancing. Finally, the monthly turnover is defined as the total value of all assets bought or sold during rebalancing divided by the portfolio value on the rebalance date. This value is bounded between 0 (no trades) and 2 (no overlapping assets pre and post rebalancing).

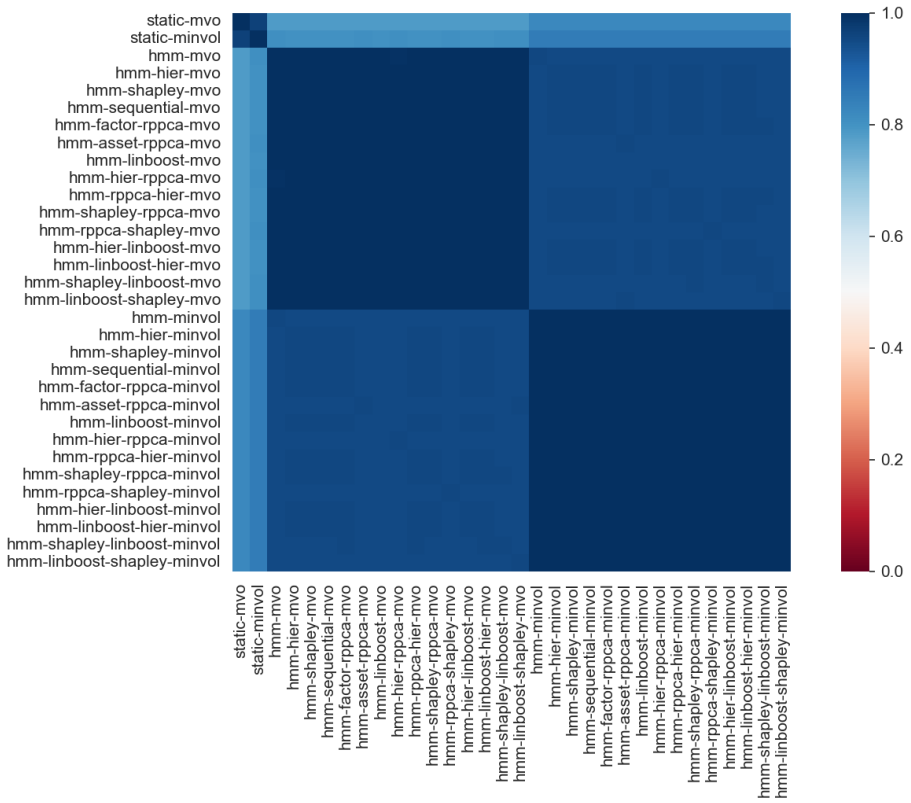

Figure 7. : Pearson correlation between the optimal weights for each model shown in Table 1, both under mean-variance optimization and minimum volatility objectives.

§8. Use Case: Qualitative Model Introspection and Interpretation

Beyond the numerical results from Section 7, another important feature of our proposed factor model is its explainability. Through the structured model framework, each of the regime detection , factor selection, and factor engineering steps can be introspected and explained. In the remainder of this section, we provide two example use cases. First, we introspect a regime-dependent model built from a 2-state GMMHMM and a Shapley Feature Selector. Next, we introspect a similar model, with a Sequential GRS Feature Selector rather than Shapley. This section does not aim to present an exhaustive set of introspection tools, rather explainers can be easily built and introduced to the pipeline based on the user's objectives.

8.1. Introspection With Shapley Feature Selection

We fit the previously described model at the end of 2022 using a 20 year lookback window. As in previous sections, Mkt-RF is used to fit the regime model, the USFM is used as the input factor model, and the asset universe is the top 200 US stocks by market cap. This results in a regime-classification that can intuitively be interpreted as a *stable* and a *volatile* state, as shown in Figure 8. As seen in Figure 9, the last few years of the training sample has been quite volatile, with all of 2022 being classified as volatile. Beyond from this, the regime classification picks up on expected events such as *Volmageddon* in

early 2018 and 2020's market crisis. This first level of introspection allows us to validate the fitted regime model against our economic intuition about market regimes, and more importantly against our investment objectives.

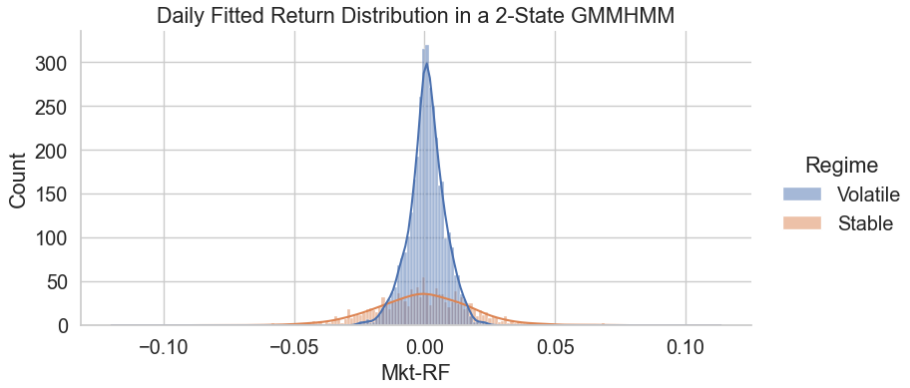

Figure 8. : Daily fitted return distribution for a 2-state GMMHMM fitted over 2003-2022.

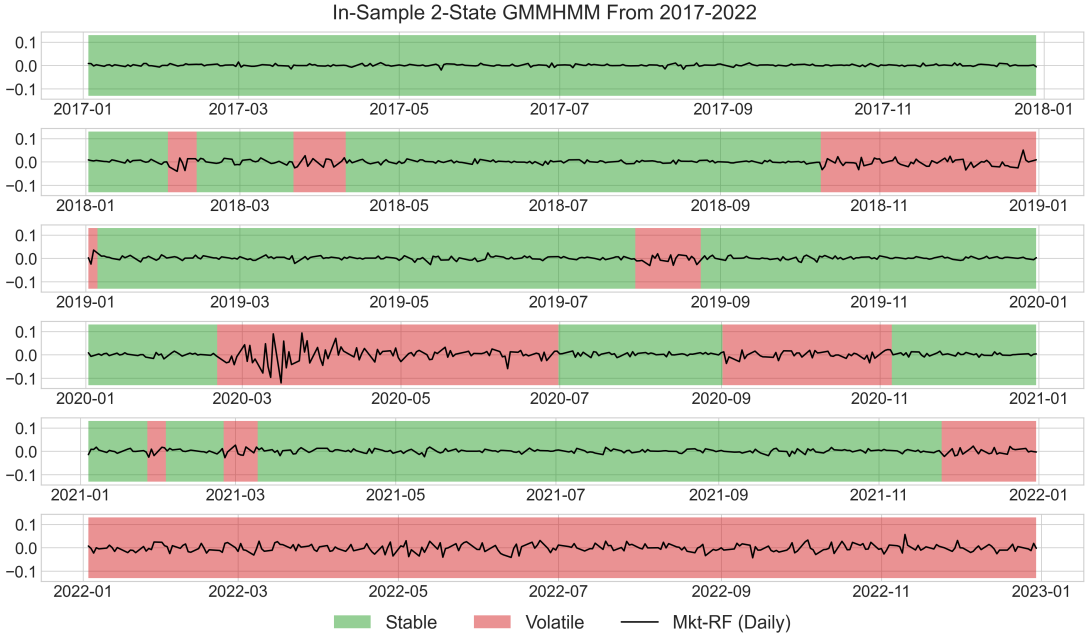

Figure 9. : In-sample regime classification plotted against market returns for the last 6 years of the training window for a 2-state GMMHMM fitted over 2003-2022.

Given this regime model, a logical next step is to investigate the selected factors for each regime, specifically where they differ. A description of each factor can be found in Appendix F. As seen in Figure 10, the majority of the factors overlap. This is expected as there are certain themes (e.g. *Low Risk* or *Market Risk*) that are almost always considered important.

Within the differing factors, the stable state contains slightly more *Value* factors, while the volatile state seems to contain slightly more *Quality*.

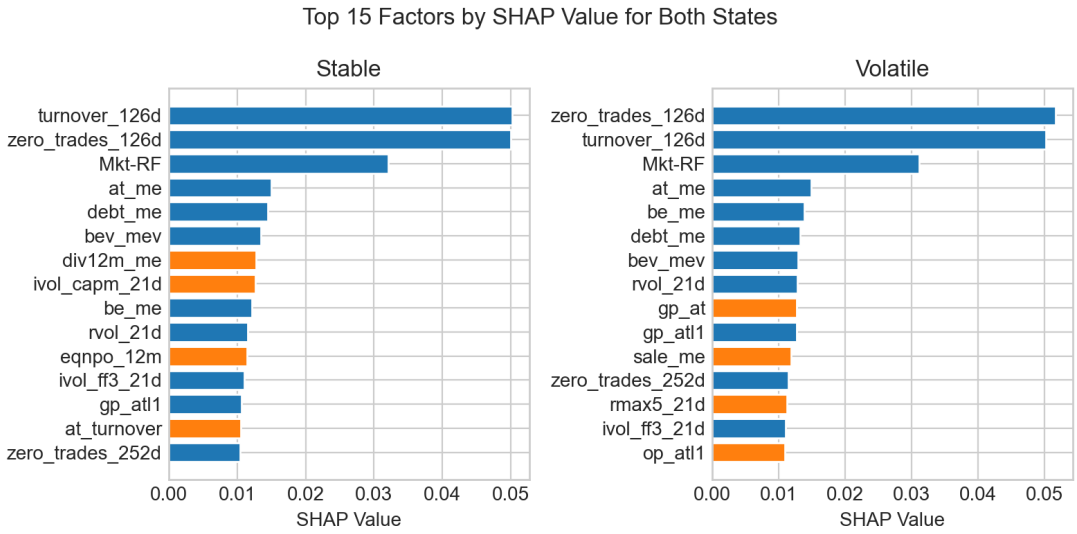

Figure 10. : Top 15 factors ranked by Shapley value for a 2-state GMMHMM fitted over 2003-2022. Overlapping factors are indicated in blue, unique factors are indicated in orange.

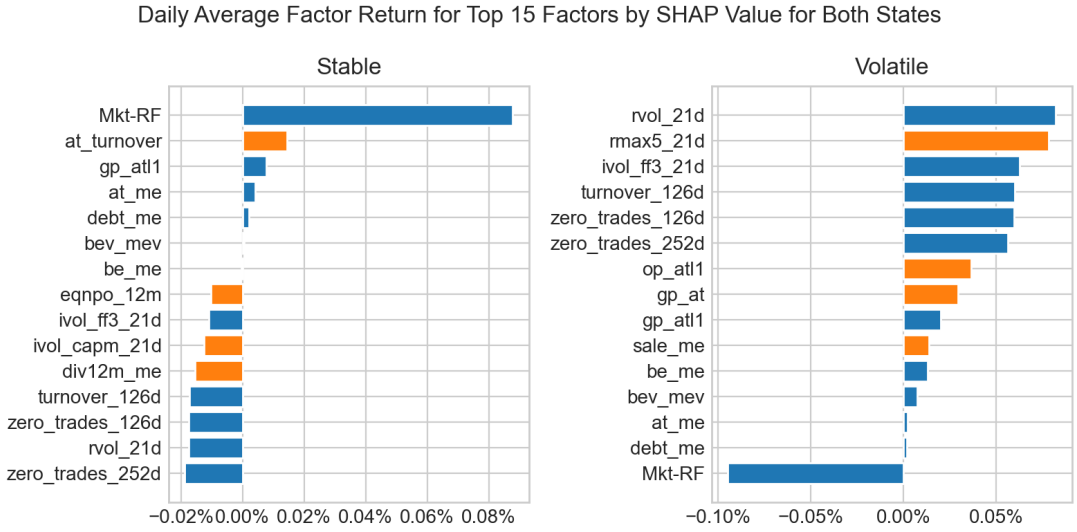

Figure 11. : Daily average returns of the top 15 factors ranked by Shapley value for a 2-state GMMHMM fitted over 2003-2022. Overlapping factors are indicated in blue, unique factors are indicated in orange.

In the stable regime, several factors are associated with Value, like debt-to-market, (`debt_me`), book-to-market (`at_me`), and dividend yield (`div12m_me`). Looking at Figure 11, *Low Leverage* factors, such as R&D-to-sales (`rd_sale`) and net debt scaled by market

equity (`netdebt_me`), show positive returns in the stable regime, while they are absent in the volatile regime. This hints at the fact that Value, Low Leverage and Quality could be partially conditional to the market environment.

In contrast, in the volatile state the emphasis is shifted towards tail-risk and volatility. In particular, Low Risk factors, e.g. betting-against-beta (`betabab_1260d`) and low idiosyncratic volatility (`ivol_ff3_21d`), help to describe the dynamics of the volatile regime. This might be explained through investors' *flight to safety* behavior during times of market distress. This idea is reinforced by the amount of Quality factors in the volatile state, mainly based on operating profits (`op_at`, `cop_atl1`,`gp_at`, `op_atl1`).

Considering all factors jointly, we observe some potential issues with this model configuration. The only factors that were selected were Mkt-Rf and a set of Value, Quality, and Low Risk factors. This would imply that 10 out of the 13 available factor themes are not relevant to explaining the asset cross-section. Such a drastic selection might not be desirable, depending on the investment objective.

Remarkably, book-to-market (HML) is the only Fama French factor that was selected through Shapley selection. Although the inclusion of HML hints at its importance in explaining returns, it experiences low average returns in both regimes. This further hints that this factor might already be priced out of the market. In addition, share turnover (`turnover_126d`) and the frequency of zero trading days (`zero_trades_126d`) are two of the most important factors and shared between both regimes. As two liquidity factors, this could highlight the importance of market frictions as a source of asset returns, independent of the current market environment.

Based on these insights, a selection of more than 15 factors or the adoption of an alternative factor selection method can be considered. Understanding how different factors interact with these market regimes can be instrumental in aligning a portfolio with investor's objectives and risk preferences.

8.2. Introspection With Sequential Selection

Next to Shapley-based factor selection, we also consider sequential factor selection through GRS, with the selected factors shown in Figure 12. A notable difference is the lack of overlapping factors between the two states, except for the market factor (Mkt-RF). Furthermore, the selected factors span a much more diversified set of themes than those selected through Shapley, as can be seen in Appendix F.

Notably, the factor directions are quite different between Shapley and GRS based factor selection. While we see many factors with an average negative return in the stable regime from Shapley, this is not the case for GRS. At the same time, there are more factors with average negative returns in the volatile regime for GRS than for Shapley. Overall, the observed differences between these two selection methods further understate that there is no one-size-fits-all solution to construct an optimal factor model.

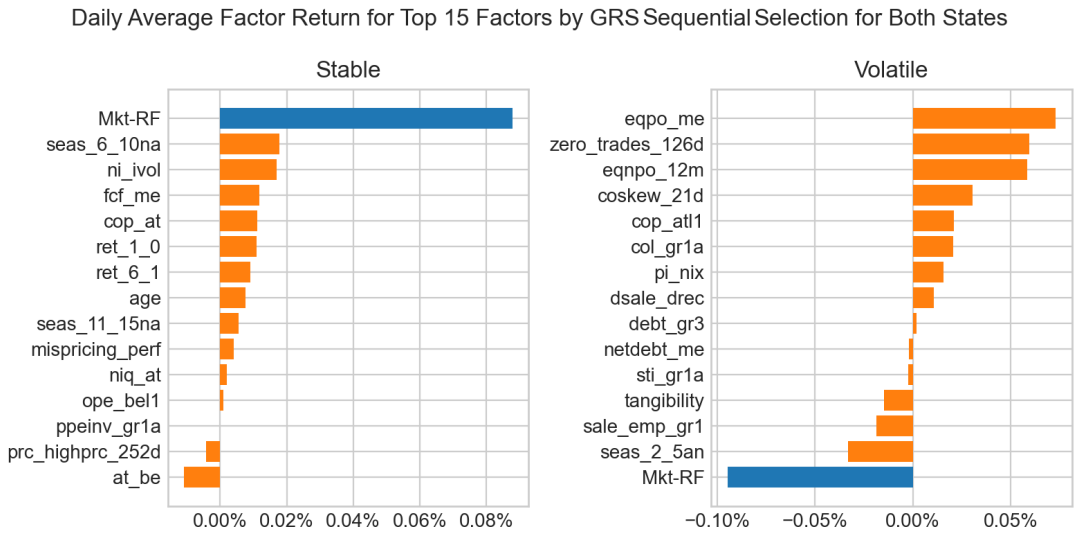

Figure 12. : Daily average returns of the top 15 factors ranked by GRS Sequential selection. Overlapping factors are indicated in blue, unique factors are indicated in orange.

§9. Discussion and Extensions

Revisiting the fundamental purpose of this paper, our goal was to offer both academics and industry professionals a more transparent, yet equally robust alternative to powerful but opaque statistical and machine learning-based factor models. In pursuit of this objective, we developed a structured factor model that integrates traditional factor modeling methods with innovative concepts from recent research. This framework is flexible and easily allows the introduction of new ideas. Finally, as shown in Section 8, the structured nature of this framework allows users to build *explainers* for every component to introspect the model end-to-end. In the remainder of this section, we will briefly discuss how the analysis of this paper could be extended and next how the proposed factor model could be improved based on what we learned.

9.1. Analysis Extensions

On the empirical side, we found mixed evidence in favour of our proposed factor model as a forecasting tool, and strong evidence in favour of our proposed factor model for portfolio optimization. A key takeaway is that there is no one-size-fits-all solution. This is both a blessing and a curse as our proposed model allows for a great level of flexibility, but subsequently needs to be fine-tuned properly.

A strong disclaimer with our results is that all model evaluation and backtesting happened over a universe of large cap, US equities. There are reasons to believe that other universes could present more interesting opportunities for our proposed factor model. First, it is well known that many equity factors are largely driven by small-caps, low-liquidity firms, or firms that experience some form of market friction. Components such as factor engineering may be more well-suited to such universes, rather than the already efficient

universe we considered.

Next, as shown by Swade, et al. [39], the US equity market is explainable by a smaller number of factors than other countries and it is one of the only markets where the country-specific factor model (USFM) outperforms the global factor model (GFM) from Jensen, et al. [26] [39]. As a result, factor selection and engineering may be more effective outside of the US.

Finally, our optimization exercise considered a very naive alpha model, namely the expected returns implied by the factor model. An interesting extension of the analysis would be to test our proposed model with different, realistic alpha models. Beyond potentially improving results, it could teach us about the source of the proposed model's performance. If all out-performance over the static model came from having better expected returns, rather than a better asset VCV estimate, then this should become apparent quickly.

9.2. Model Extensions

Most notably, we see a clear path ahead to extend and improve the proposed model framework. A first order improvement could come from improved regime detection models. Hirsa, et al. [23] recently proposed Robust Rolling Regime Detection, which promises to be highly stable under re-training, a key problem with HMMs. Next to that, new data and models also open up new ways to think about regime detection. For instance, one could look towards options-markets to construct probability density functions of the market [38] [29]. These distributions could then be fed into an algorithm such as Wasserstein K-Means which is able to detect regimes over probability distributions [24].

A second improvement related to the regime-conditioning is a more sophisticated factor blending method. The current method, while intuitive to understand and computationally efficient, is somewhat naive. Concretely, we would suggest to look beyond a single period to match the model closer to the investment horizon. On the factor model side, this could for instance happen by using modified regime transition probabilities that look N days ahead. On the optimization side, the Markov Chain underlying HMM could be exploited to maximize objectives under stochastic programming [1]. Under some assumptions, stochastic programming can be performed for regime models other than HMMs as shown during a previous project [11]. Conceptually, stochastic programming is just a simulation tool (in this case HMM) combined with a numerical optimizer. More generally, this framework could lend itself well to advanced techniques such as reinforcement learning. Although, that would no longer be in line with our interpretability objective.

Next to new or more advanced features, the model can be improved through enhanced robustness. A logical improvement would be to incorporate some form of covariance shrinkage on the factor VCVs to limit noise sensitivity [40]. Additionally, the trade-off between modeling systematic and idiosyncratic asset behavior can be tackled more explicitly. For instance, it is well known that idiosyncratic volatility is mean-reverting [4]. Similarly, idiosyncratic returns (i.e. α) may be overestimated in the current model. Incorporating such ideas can aid in both the conceptual and numerical robustness of the proposed model.

§10. Conclusion

As highlighted in the introduction, our research sought to extend the existing knowledge in regime-dependent portfolio construction and factor models across three pivotal areas. First, whereas most papers have approached regime-dependent portfolio construction as an asset or index allocation problem, this paper applied and evaluated models over a large equity cross-section. We found moderate evidence on the performance of our model from an equity risk and return forecasting perspective, and positive evidence from a portfolio optimization perspective. We recognize that there is no one-size-fits-all solution, and our model addresses this by giving users the flexibility to fine-tune the pipeline towards their objectives.

Second, this paper considered the joint problem of regime-modeling, compressing the factor zoo, and finding new factors. Although each of these have been researched independently, to our knowledge they have not yet been considered jointly. We only found mixed evidence in favor of factor engineering for US, large-cap equities. However, we found strong positive evidence on the impact of regime-conditioning on factor models and find that factor selection can be applied successfully within regimes. Different factor selection methods offer similar results, presenting a high level of flexibility.

Finally, this paper proposes a structured and explainable framework as an alternative to recent developments in powerful, but opaque ML-based latent factor models. As shown through an example use case, our model can be interpreted end-to-end, giving valuable insights into the dynamics of factors and regimes. Furthermore, due to the structured nature of the model, new introspection tools can be easily incorporated to satisfy the user's objectives.

References

[1] G. I. Bae, W. C. Kim, and J. M. Mulvey. Dynamic asset allocation for varied financial markets under regime switching framework. *European Journal of Operational Research*, 234(2):450–458, 2014.

[2] L. E. Baum. An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. In O. Shisha, editor, *Inequalities III: Proceedings of the Third Symposium on Inequalities*, pages 1–8, University of California, Los Angeles, 1972. Academic Press.

[3] P. Bilokon, A. Jacquier, and C. McIndoe. Market regime classification with signatures. *arXiv preprint arXiv:2107.00066*, 2021.

[4] S. Bozhkov, H. Lee, U. Sivarajah, S. Despoudi, and M. Nandy. Idiosyncratic risk and the cross-section of stock returns: the role of mean-reverting idiosyncratic volatility. *Annals of Operations Research*, 294:419–452, 2020.

[5] M. Cerliani. linear-tree, 2021. URL https://github.com/cerlymarco/linear-tree.

[6] R. Clarke, H. De Silva, and S. Thorley. Portfolio constraints and the fundamental law of active management. *Financial Analysts Journal*, 58(5):48–66, 2002.

[7] G. Coqueret. Diversified minimum-variance portfolios. *Annals of Finance*, 11(2): 221–241, 2015.

[8] G. Costa and R. H. Kwon. Risk parity portfolio optimization under a markov regime-switching framework. *Quantitative Finance*, 19(3):453–471, 2019.

[9] G. Costa and R. H. Kwon. A regime-switching factor model for mean-variance optimization. *Journal of Risk*, 2020.

[10] M. L. De Prado. Building diversified portfolios that outperform out of sample. *The Journal of Portfolio Management*, 42(4):59–69, 2016.

[11] Y. A. D'hondt, M. M. Di Venti, R. Rishi, and J. Walker. MarketMoodRing, July 2023. URL https://github.com/yvesdhondt/MarketMoodRing.

[12] Y. A. D'hondt, A. Gulati, M. M. Di Venti, R. Rishi, and J. Walker. AFP Online Appendix, 2024. URL https://github.com/yvesdhondt/AFP-Online-Appendix/tree/main.

[13] Y. Duan, L. Wang, Q. Zhang, and J. Li. Factorvae: A probabilistic dynamic factor model based on variational autoencoder for predicting cross-sectional stock returns. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4468–4476, 2022.

[14] E. F. Fama and K. R. French. The value premium. *The Review of Asset Pricing Studies*, 11(1):105–121, 2021.

[15] K. R. French. Fama/french 3 factors [daily], 2023. URL https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html. Retrieved from Kenneth R. French Data Library.

[16] N. Gârleanu and L. H. Pedersen. Dynamic trading with predictable returns and transaction costs. *The Journal of Finance*, 68(6):2309–2340, 2013.

[17] M. R. Gibbons, S. A. Ross, and J. Shanken. A test of the efficiency of a given portfolio. *Econometrica: Journal of the Econometric Society*, pages 1121–1152, 1989.

[18] R. C. Green and B. Hollifield. When will mean-variance efficient portfolios be well diversified? *The Journal of Finance*, 47(5):1785–1809, 1992.

[19] M. Guidolin and A. Timmermann. Asset allocation under multivariate regime switching. *Journal of Economic Dynamics and Control*, 31(11):3503–3544, 2007.
[20] M. Guidolin and A. Timmermann. International asset allocation under regime switching, skew, and kurtosis preferences. *The Review of Financial Studies*, 21(2):889–935, 02 2008.
[21] J. D. Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2):357–384, 1989. ISSN 00129682, 14680262.
[22] Z. L. He. Comparing asset pricing models: Distance-based metrics and bayesian interpretations. *arXiv preprint arXiv:1803.01389*, 2018.
[23] A. Hirsa, S. Xu, and S. Malhotra. Robust rolling regime detection (r2-rd): A data-driven perspective of financial markets, Feb. 16, 2024.
[24] B. Horvath, Z. Issa, and A. Muguruza. Clustering market regimes using the wasserstein distance. *arXiv preprint arXiv:2110.11848*, 2021.
[25] I. Ilic, B. Görgülü, M. Cevik, and M. G. Baydoğan. Explainable boosted linear regression for time series forecasting. *Pattern Recognition*, 120:108144, 2021.
[26] T. I. Jensen, B. T. Kelly, and L. H. Pedersen. Is there a replication crisis in finance? *The Journal of Finance*, 78(5):2465–2518, 2023.
[27] D. Jurafsky and J. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, volume 2. 02 2008.
[28] M. J. Kamstra and R. Shi. A note on the grs test. *Available at SSRN 3775089*, 2020.
[29] Y. Kitsul and J. H. Wright. The economics of options-implied inflation probability density functions. *Journal of Financial Economics*, 110(3):696–711, 2013.
[30] O. Ledoit and M. Wolf. Honey, i shrunk the sample covariance matrix. *UPF economics and business working paper*, (691), 2003.
[31] M. Lettau and M. Pelger. Estimating latent asset-pricing factors. *Journal of Econometrics*, 218(1):1–31, 2020.
[32] M. Levy and Y. Ritov. Mean–variance efficient portfolios with many assets: 50% short. *Quantitative Finance*, 11(10):1461–1471, 2011.
[33] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.
[34] J. Pfitzinger, N. Katzke, et al. A constrained hierarchical risk parity algorithm with cluster-based capital allocation. *Stellenbosch University, Department of Economics*, 2019.
[35] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.
[36] T. Raffinot. The hierarchical equal risk contribution portfolio. *Available at SSRN 3237540*, 2018.
[37] L. S. Shapley. A value for n-person games. *Princeton University Press*, 1953.
[38] D. Shimko. Bounds of probability. *Risk*, 6(4):33–37, 1993.
[39] A. Swade, M. X. Hanauer, H. Lohre, and D. Blitz. Factor zoo (. zip). *Available at SSRN*, 2023.

[40] J. Theiler. The incredible shrinking covariance estimator. In *Automatic target recognition XXII*, volume 8391, pages 225–236. SPIE, 2012.

[41] J. Tu. Is regime switching in stock returns important in portfolio decisions? *Management Science*, 56(7):1198–1215, 2010.

[42] C. M. Turner, R. Startz, and C. R. Nelson. A markov model of heteroskedasticity, risk, and learning in the stock market. *Journal of Financial Economics*, 25(1):3–22, 1989.

[43] Z. Wei, A. Rao, B. Dai, and D. Lin. Hirevae: An online and adaptive factor model based on hierarchical and regime-switch vae. *arXiv preprint arXiv:2306.02848*, 2023.

[44] Wharton Research Data Services. WRDS. wrds.wharton.upenn.edu. Accessed 2024-02-18.

§A. Hidden Markov Model Description

Following Jurafsky, et al. [27], a Hidden Markov Model is defined in detail by:

1. A sequence of **observations** from a (multivariate) time series $\{r_t\}_{t=1}^T$ (returns in the context of this paper)
2. A set of K **latent states** $Q_1, Q_2, ..., Q_K$ (market regimes in the context of this paper)
3. A transition probability matrix $A_{ij} = P(S_t = Q_j | S_{t-1} = Q_i)$
4. A sequence of **observational likelihoods**, called emission probabilities $B = b_i(r_t)$, that each express the probability of observation r_t being generated from state i
5. An **initial probability distribution** over the states $\pi \in \mathbb{R}_+^K$ such that $\pi^T \mathbf{1} = 1$

The first order Hidden Markov Model used in this paper makes use of two additional assumptions:

1. **Markov Assumption**: $P(s_t | s_1, \ldots, s_{t-1}) = P(s_t | s_{t-1})$
2. **Output Independence**: $P(r_t | s_1, \ldots, s_T, r_1, \ldots, r_T) = P(r_t, s_t)$

Rabiner [35] characterises HMMs by three fundamental problems:

1. **Likelihood estimation**: Given an HMM, transition matrix and emission probabilities (A, B), and an observation sequence $O = \{r_t\}_{t=1}^T$, determine the likelihood $P(O|A, B)$
2. **Decoding**: Given an HMM, (A, B), and observation sequence $\{r_t\}_{t=1}^T$, determine the best hidden state sequence $\{s_t\}_{t=1}^T$
3. **Learning**: Given observation sequence $\{r_t\}_{t=1}^T$, and the set of states in the HMM, determine the best parameters $\theta = (A, B)$

The learning and decoding problems are of primary interest for latent regime detection. By prescribing the number of latent states and the emission distribution, the set of parameters, $\theta = (A, B)$, can be learned through the forward-backward algorithm, also known as the Baum-Welch algorithm [2], a special case of the Expectation-Maximization algorithm.

§B. Note on the GRS Statistic

Kamstra, et al. provide a detailed note and proof of the correct GRS statistic [28]. Consider a linear factor model fitted over T periods of time, for L factors, and N test assets:

$$r^e_{i,t} = \alpha_i + \sum_{j=1}^{L} \beta_{i,j} f_{j,t} + \epsilon_{i,t}$$

Here, $r^e_{i,t}$ are the excess returns of test-asset i at time t, $f_{j,t}$ are the factor returns for factor j at time t, and $\epsilon_{i,t}$ is the regression residual. For this model, the GRS statistic $\tilde{W}$ is defined as follows:

$$\tilde{W} \equiv \frac{T(T-N-L)}{N(T-L-1)} \left(1 + \bar{f}^T \tilde{\Omega}^{-1} \bar{f}\right) \hat{\alpha}^T \hat{\Sigma}^{-1} \hat{\alpha}$$

Here, $\bar{f}$ is the vector of expected factor returns, and α is the vector of regression alphas from the linear factor model. Now the point of caution lies in $\tilde{\Omega}$ and $\hat{\Sigma}$. $\tilde{\Omega}$ is the covariance matrix of the factor returns and should be estimated with 0 degrees of freedom, i.e. as if it was a population covariance matrix. A common mistake is to calculate this covariance matrix with 1 degree of freedom which leads to an ill-distributed test statistic [28]. Finally, $\hat{\Sigma}$ is the covariance matrix of the regression residuals and should be estimated with $L+1$ degrees of freedom.

Kamstra, et al. go on to show that if and only if the GRS test statistic is calculated under this specification, the test follows an F distribution [28]:

$$\tilde{W} \sim F_{N,T-N-L}$$

§C. Hierarchical Risk Parity - Distance Metric

de Prado proposes to exploit the network structure between the assets by finding a minimum spanning tree that describes the majority of the distance structure between the assets. To achieve this, we can encode the distance between any two assets a_i and a_j as a function of the correlation, $\rho_{i,j}$ between those assets [10]:

$$d(a_i, a_j) = \sqrt{\frac{1}{2}(1 - \rho_{i,j})}$$

We can calculate this distance metric for each pair of assets, giving rise to a distance matrix $D(U)$ where $D_{i,j} = d(a_i, a_j)$. This distance metric is then extended into another distance metric which encodes not only the distance between each pair of assets in isolation, but rather between each pair of assets taking into account all other assets. This is achieved by considering the Euclidean distance between each pair of columns of D [10]:

$$\tilde{d}(a_i, a_j) = \sqrt{\sum_{k=1}^{N}(D_{k,i} - D_{k,j})^2}$$

We can calculate this distance metric for each pair of assets, giving rise to a distance matrix $\tilde{D}(U)$ where $\tilde{D}_{i,j} = \tilde{d}(a_i, a_j)$ [10]. This matrix will also be symmetrical and requires the estimation of $\frac{N(N+1)}{2}$ elements. The innovation arises from now applying agglomerative clustering over this distance matrix to find a unique dendrogram that encodes the distance relationships of the entire network. The specific linkage algorithm to use can be viewed as a hyperparameter, with common options being single, complete, average, and Ward linkage. As a dendrogram is both a minimum spanning subtree of the network as well as a strict binary tree, it is entirely defined by $2N - 1$ nodes and $2N - 2$ edges for a total of $4N - 3$ elements. Comparing this to an asset VCV matrix, we have moved from a representation complexity of $O(N^2)$ to $O(N)$.

§D. Blending N Regime-Conditional Factor Models

Consider again the set-up from Section 4.4. Let $f_{t,i} \sim N(\mu_{f_i}, F_i)$ where μ_{f_i} are the expected factor returns in state i. Next, allow for different factor sets for each regime, i.e. $f_{t,i} \in \mathbb{R}^{L_i}$ where the number of factors in state i, L_i, and the set of factors is determined by the factor selection and engineering procedures outlined above. Under the model specification from Eq. (3.2), we can now blend the expected return vectors and VCVs for any N regimes. Let the asset expected excess return vector within a fixed state j be given by:

$$\mu_j = \alpha_j + K_j^T \mu_{f_j} \tag{D.1}$$

Where the intercept (α_j) and factor loadings K_j can be estimated for each asset through OLS.

Then for an N-state model, the asset expected excess return vector and VCV conditional on being in state i today are given by:

$$\hat{\mu}_i^e = \sum_{j=1}^{N} \gamma_{i,j} \mu_j \tag{D.2}$$

$$\begin{aligned} \hat{\Sigma}_i = &\sum_{j=1}^{N} \gamma_{i,j}(K_j^T F_j K_j + D_j) + \sum_{j=1}^{N} \gamma_{i,j}(1-\gamma_{i,j})\mu_j \mu_j^T \\ &- \sum_{k,j,k\neq j}^{N} \gamma_{i,k}\gamma_{i,j}\mu_k \mu_j^T \end{aligned} \tag{D.3}$$

§E. Factor Model Evaluation Results

E.1. Wasserstein Distance

Wasserstein Distance						
	1 Regime	2 Regimes				
Validation Window	Static USFM	GMMHMM (None)	GMMHMM (GRS)	GMMHMM (Shapley)	GMMHMM (Hierarchical)	GMMHMM (RP-PCA)
20 Days	0.270 (0.036)	0.264 (0.071)	0.264 (0.071)	0.264 (0.071)	0.265 (0.071)	0.264 (0.071)
40 Days	0.255 (0.053)	0.252 (0.083)	0.251 (0.083)	0.251 (0.083)	0.252 (0.083)	0.252 (0.083)
62 Days	0.239 (0.049)	0.237 (0.080)	0.237 (0.080)	0.237 (0.080)	0.238 (0.080)	0.238 (0.080)
250 Days	0.201 (0.039)	0.200 (0.074)	0.200 (0.074)	0.200 (0.074)	0.201 (0.074)	0.200 (0.074)
500 Days	0.190 (0.036)	0.194 (0.073)	0.194 (0.074)	0.193 (0.074)	0.195 (0.073)	0.194 (0.073)
750 Days	0.185 (0.036)	0.191 (0.073)	0.191 (0.073)	0.191 (0.073)	0.192 (0.073)	0.192 (0.073)

Table 5.: Mean Wasserstein distance for different selection methods and different validation window sizes, using 3,750 training days and a 62 day stride. Standard deviations are denoted in brackets ().

Wasserstein Distance						
	1 Regime	3 Regimes				
Validation Window	Static USFM	GMMHMM (None)	GMMHMM (GRS)	GMMHMM (Shapley)	GMMHMM (Hierarchical)	GMMHMM (RP-PCA)
20 Days	0.270 (0.036)	0.262 (0.075)	0.271 (0.078)	0.263 (0.079)	0.265 (0.075)	0.261 (0.069)
40 Days	0.255 (0.053)	0.250 (0.085)	0.258 (0.086)	0.250 (0.088)	0.252 (0.084)	0.248 (0.080)
62 Days	0.239 (0.049)	0.236 (0.083)	0.244 (0.084)	0.236 (0.086)	0.239 (0.081)	0.234 (0.077)
250 Days	0.201 (0.039)	0.206 (0.081)	0.209 (0.082)	0.206 (0.086)	0.207 (0.078)	0.207 (0.075)
500 Days	0.190 (0.036)	0.200 (0.082)	0.204 (0.085)	0.199 (0.088)	0.201 (0.080)	0.201 (0.075)
750 Days	0.185 (0.036)	0.191 (0.069)	0.194 (0.070)	0.188 (0.074)	0.192 (0.066)	0.196 (0.075)

Table 6.: Mean Wasserstein distance for different selection methods and different validation window sizes, using 3,750 training days and a 62 day stride. Standard deviations are denoted in brackets ().

Wasserstein Distance				
	1 Regime	2 Regimes		
Validation Window	Static USFM	GMMHMM (None)	GMMHMM (Linear Boosting)	GMMHMM (RP-PCA)
20 Days	0.270 (0.036)	0.264 (0.071)	0.264 (0.071)	0.265 (0.071)
40 Days	0.255 (0.053)	0.252 (0.083)	0.252 (0.083)	0.252 (0.083)
62 Days	0.239 (0.049)	0.237 (0.080)	0.238 (0.080)	0.238 (0.080)
250 Days	0.201 (0.039)	0.200 (0.074)	0.200 (0.074)	0.201 (0.074)
500 Days	0.190 (0.036)	0.194 (0.073)	0.194 (0.073)	0.195 (0.073)
750 Days	0.185 (0.036)	0.191 (0.073)	0.191 (0.073)	0.193 (0.073)

Table 7.: Mean Wasserstein distance for different engineering methods and validation window sizes, using 3,750 training days and a 62 day stride. Standard deviations are denoted in brackets ().

Wasserstein Distance				
	1 Regime	3 Regimes		
Validation Window	Static USFM	GMMHMM (None)	GMMHMM (Linear Boosting)	GMMHMM (RP-PCA)
20 Days	0.270 (0.036)	0.262 (0.075)	0.261 (0.073)	0.272 (0.084)
40 Days	0.255 (0.053)	0.250 (0.085)	0.248 (0.082)	0.261 (0.092)
62 Days	0.239 (0.049)	0.236 (0.083)	0.235 (0.080)	0.247 (0.091)
250 Days	0.201 (0.039)	0.206 (0.081)	0.202 (0.080)	0.213 (0.089)
500 Days	0.190 (0.036)	0.200 (0.082)	0.195 (0.081)	0.209 (0.092)
750 Days	0.185 (0.036)	0.191 (0.069)	0.182 (0.064)	0.200 (0.081)

Table 8.: Mean Wasserstein distance for different engineering methods and different validation window sizes, using 3,750 training days and a 62 day stride. Standard deviations are denoted in brackets ().

E.2. Pearson Correlation on Returns

Pearson Correlation Between Expected and Realized Returns						
	1 Regime	2 Regimes				
Validation Window	Static USFM	GMMHMM (None)	GMMHMM (GRS)	GMMHMM (Shapley)	GMMHMM (Hierarchical)	GMMHMM (RP-PCA)
20 Days	-0.014 (0.138)	0.051 (0.192)	0.052 (0.192)	0.052 (0.192)	0.053 (0.192)	0.052 (0.192)
40 Days	0.019 (0.136)	0.034 (0.187)	0.034 (0.187)	0.034 (0.187)	0.034 (0.187)	0.034 (0.187)
62 Days	0.026 (0.148)	0.053 (0.190)	0.053 (0.191)	0.053 (0.190)	0.053 (0.190)	0.053 (0.190)
250 Days	0.035 (0.172)	0.075 (0.200)	0.075 (0.200)	0.076 (0.199)	0.075 (0.200)	0.075 (0.199)
500 Days	0.033 (0.168)	0.092 (0.196)	0.091 (0.196)	0.092 (0.195)	0.093 (0.196)	0.092 (0.196)
750 Days	0.026 (0.151)	0.119 (0.190)	0.118 (0.191)	0.120 (0.189)	0.121 (0.189)	0.119 (0.190)

Table 9.: Mean Pearson correlation between expected and realized returns for different selection methods and for different validation window sizes, using 3,750 training days and a 62 day stride. Standard deviations are denoted in brackets ().

Pearson Correlation Between Expected and Realized Returns						
	1 Regime	3 Regimes				
Validation Window	Static USFM	GMMHMM (None)	GMMHMM (GRS)	GMMHMM (Shapley)	GMMHMM (Hierarchical)	GMMHMM (RP-PCA)
20 Days	-0.014 (0.138)	0.019 (0.173)	0.037 (0.181)	0.039 (0.176)	0.037 (0.166)	0.050 (0.171)
40 Days	0.019 (0.136)	0.012 (0.182)	0.032 (0.182)	0.021 (0.185)	0.023 (0.182)	0.042 (0.184)
62 Days	0.026 (0.148)	0.035 (0.179)	0.071 (0.181)	0.048 (0.179)	0.044 (0.176)	0.059 (0.183)
250 Days	0.035 (0.172)	0.064 (0.178)	0.052 (0.177)	0.062 (0.187)	0.062 (0.183)	0.070 (0.188)
500 Days	0.033 (0.168)	0.108 (0.197)	0.079 (0.190)	0.112 (0.175)	0.085 (0.175)	0.101 (0.205)
750 Days	0.026 (0.151)	0.143 (0.172)	0.121 (0.172)	0.159 (0.161)	0.128 (0.158)	0.145 (0.182)

Table 10.: Mean Pearson correlation between expected and realized returns for different selection methods and for different validation window sizes, using 3,750 training days and a 62 day stride. Standard deviations are denoted in brackets ().

Pearson Correlation Between Expected and Realized Returns				
	1 Regime	2 Regimes		
Validation Window	Static USFM	GMMHMM (None)	GMMHMM (Linear Boosting)	GMMHMM (RP-PCA)
20 Days	-0.014 (0.138)	0.051 (0.192)	0.051 (0.193)	0.052 (0.193)
40 Days	0.019 (0.136)	0.034 (0.187)	0.033 (0.187)	0.034 (0.187)
62 Days	0.026 (0.148)	0.053 (0.190)	0.053 (0.190)	0.053 (0.191)
250 Days	0.035 (0.172)	0.075 (0.200)	0.075 (0.200)	0.075 (0.200)
500 Days	0.033 (0.168)	0.092 (0.196)	0.092 (0.196)	0.091 (0.196)
750 Days	0.026 (0.151)	0.119 (0.190)	0.120 (0.190)	0.119 (0.190)

Table 11.: Mean Pearson correlation between expected and realized returns for different engineering methods and for different validation window sizes, using 3,750 training days and a 62 day stride. Standard deviations are denoted in brackets ().

Pearson Correlation Between Expected and Realized Returns				
	1 Regime	3 Regimes		
Validation Window	Static USFM	GMMHMM (None)	GMMHMM (Linear Boosting)	GMMHMM (RP-PCA)
20 Days	-0.014 (0.138)	0.019 (0.173)	0.028 (0.174)	0.017 (0.197)
40 Days	0.019 (0.136)	0.012 (0.182)	0.019 (0.185)	-0.001 (0.189)
62 Days	0.026 (0.148)	0.035 (0.179)	0.044 (0.176)	0.033 (0.188)
250 Days	0.035 (0.172)	0.064 (0.178)	0.082 (0.176)	0.055 (0.184)
500 Days	0.033 (0.168)	0.108 (0.197)	0.101 (0.176)	0.091 (0.184)
750 Days	0.026 (0.151)	0.143 (0.172)	0.137 (0.157)	0.126 (0.177)

Table 12.: Mean Pearson correlation between expected and realized returns for different engineering methods and for different validation window sizes, using 3,750 training days and a 62 day stride. Standard deviations are denoted in brackets ().

§F. Model Introspection: Selected Factors

Factor	Description	Theme	Sign
Mkt-RF*ˆ	Market Excess Returns - Sharpe(1964)	N/A	1
debt_gr3ˆ	Growth in book debt (3 years) - Lyandres, Sun, and Zhang (2008)	Debt Insurance	-1
col_gr1aˆ	Change in current operating liabilities - Richardson et al. (2005)	Investment	-1
ppeinv_gr1aˆ	Change PPE and Inventory - Lyandres et al. (2008)	Investment	-1
at_me*	Total Assets scaled by Market Equity – Fama and French (1992)	Value	1
be_me*	Book-to-market equity - Rosenberg, Reid, and Lanstein (1985)	Value	1
bev_mev*	Book Enterprise Value scaled by Market Equity Value – Penman	Value	1
debt_me*	Debt-to-market – Bhandari (1988)	Value	1
div_12m_me*	Dividend yield – Litzenberger and Ramaswamy (1979)	Value	1
eqnpo_12m*ˆ	Equity net payout – Daniel and Titman (2006)	Value	1
eqpo_meˆ	Payout yield - Boudoukh et al. (2007)	Value	1
fcf_meˆ	Free cash flow-to-price - Lakonishok et al. (1994)	Value	1
sale_me*	Sales-to-market - Barbee, Mukherji, and Raines (1996)	Value	1
ageˆ	Firm age - Jiang, Lee, and Zhang (2005)	Low Leverage	-1
at_beˆ	Book leverage - Fama and French (1992)	Low Leverage	-1
netdebt_meˆ	Net debt-to-price - Penman, Richardson, and Tuna (2007)	Low Leverage	-1
ni_ivolˆ	Earnings Volatility -Francis et al. (2004)	Low Leverage	1
tangibilityˆ	Asset tangibility - Hahn and Lee (2009)	Low Leverage	1
zero_trades_126d*ˆ	Zero trades with turnover as a tiebreak (6 months) – Liu (2006)	Low Risk	1
ivol_capm_21d*	Idiosyncratic volatility over a 21-day period from CAPM – Francis et Al. (2004)	Low Risk	-1
ivol_ff3_21d*	Idiosyncratic volatility from the Fama-French 3-factor model – Ang et al. (2006)	Low Risk	-1
rmax5_21d*	Highest 5 days of return - Bali, Brown, Murray and Tang (2017)	Low Risk	-1
rvol_21d*	Return volatility – Ang, Hodrick, et Al.	Low Risk	-1
turnover_126d*	Share turnover 6 months – Datar, Naik, and Radcliffe (1998)	Low Risk	-1
zero_trades_252d*	Zero trades with turnover as a tiebreak (12 months) – Liu (2006)	Low Risk	1
prc_highprc_252dˆ	Current price to high price over last year - George and Hwang (2004)	Momentum	1
ret_6_1ˆ	Price momentum t-6 to t-1 - Jegadeesh and Titman (1993)	Momentum	1
dsale_drecˆ	Change sales minus change receivables - Abarbanell and Bushee (1998)	Profit Growth	1
sale_emp_gr1ˆ	Labor force efficiency - Abarbanell and Bushee (1998)	Profit Growth	1
ope_bel1ˆ	Operating profits-to-lagged book equity - Fama and French (2015)	Profitability	1
cop_atˆ	Cash-based operating profits-to- book assets	Quality	1
cop_atl1ˆ	Cash-based operating profits-to-lagged book assets - Nikolaev et al. (2016)	Quality	1
mispricing_perfˆ	Mispricing factor: Performance - Stambaugh and Yuan (2017)	Quality	1
niq_atˆ	Quarterly return on assets - Balakrishnan, Bartov, and Faurel (2010)	Quality	1
at_turnover*	Capital turnover - Haugen and Baker (1996)	Quality	1
gp_at*	Gross Profits-to-Assets – Novy-Marx (2013)	Quality	1
gp_atl1*	Gross Profit scaled by lagged Assets – Novy-Marx (2013)	Quality	1
op_atl1*	Operating profits-to-lagged book assets – Ball et Al. (2016)	Quality	1
coskew_21dˆ	Coskewness - Harvey and Siddique (2000)	Seasonality	-1
pi_nixˆ	Taxable income-to-book income - Lev and Nissim (2004)	Seasonality	1
seas_11_15naˆ	Years 11-15 lagged returns, non-annual - Heston and Sadka (2008)	Seasonality	-1
seas_2_5anˆ	Years 2-5 lagged returns, annual - Heston and Sadka (2008)	Seasonality	1
seas_6_10anˆ	Years 6-10 lagged returns, annual - Heston and Sadka (2008)	Seasonality	1
sti_gr1aˆ	Change in short-term investments - Richardson et al. (2005)	Seasonality	1
ret_1_0ˆ	Short-term reversal - Jegadeesh (1990)	Short Term Reversal	-1

Table 13.: Set of factors selected in the example use cases of Section 8. Shapley-selected factors are indicated with a star (*) and GRS-selected factors with a hat (ˆ). All factors are provided in the online resources to Jensen, et al. [26]. Sign indicates whether the high tercile of the characteristic is long (1) or short (-1) in the long-short tercile portfolio constructing the factor return.