Factor games: May the p-values be ever in your favor*

Hugo Finizola Stellet†, Fernando Tassinari Moraes‡

This version: April 2024

Abstract

This paper introduces a novel, time-varying framework for high-dimensional factor-based asset pricing models. It leverages shrinkage techniques on regressions spanning pricing anomalies to identify statistically significant factors from a vast pool, akin to a financial Hunger Games – the Factor Games – where may the odds (p-values) be ever in their favor. The framework emphasizes sparsity, proposing methods to select a limited number of impactful factors and outperform a stricter benchmark that incorporates the Fama-French 3-factor model proposed methodology - all while avoiding look-ahead bias. Recognizing the implicit sparsity assumption in traditional models, the framework explicitly considers similar scarcity during factor selection. We apply the proposed framework to a large set of factors and various time periods, demonstrating that simple techniques can yield interesting results when applied with proper methodology. Overall, this paper provides valuable tools for researchers and practitioners, offering guidance for pricing factor selection and advocating for sparsity.

Keywords— Factor investing, Shrinkage penalization, Time-varying asset pricing, Statistical significance

May God forgive those bad people.

Adriano "Imperador" Ribeiro

*I am indebted to Marcelo Medeiros for advising this project. I must also thank Chuanping Sun for his help with the data. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

†PhD candidate at PUC-Rio (hfstellet@poli.ufrj.br)

‡Insper - Institute of Education and Research (fernandotm@al.insper.edu.br)

Contents

1 Introduction

In the dystopian nation of Panem, the lingering shadow of a past rebellion weighs heavily upon its twelve districts. The Hunger Games, a brutal televised event orchestrated by the extravagant Capitol, stands as a grim punishment for the districts' past defiance. Each year, these Games force children, known as tributes, to battle to the death in a savage test of survival, pitting tributes from each district against one another until only one emerges victorious from the carnage.

Imagine a similar contest, not with human lives at stake, but with the alleged anomaly factors that influence how assets are priced. These factors, much like the tributes, represent various influences, some well-established, others more controversial, on asset returns. This financial rendition of the Hunger Games employs a series of spanning regressions, with a shrinkage technique serving as our arena master, compelling these factors to confront each other directly. In this framework, resembling the Games' arena, each anomaly's return undergoes regression against the returns of its counterparts. Here, the shrinkage technique meticulously tests if any anomaly could have their returns explained by a cunning combination of the others.

However, our objective is not the literal elimination of infants but rather the identification of statistically significant factors. Just as a tribute's survival in the Games hinges on its strength and tactics, a factor's explanatory prowess is determined by the statistical significance of its intercept in the spanning regression. Factors with non-significant - high p-values - intercepts, like tributes who fall early in the Games, are deemed unable to explain asset returns independently when their peers are already accounted for. Conversely, factors boasting statistically significant intercepts, akin to the Games' victor, have demonstrated their unique impact on asset returns, distinguishing themselves from the competition and asserting their significance in this economic arena. Our research endeavors to unveil these triumphant factors, the true heavyweights in the quest to explain asset returns. May the p-values be ever in your favor!

The quest for factors that potentially explain aspects of the cross-section of expected returns has spawned (i)numerous contenders in the literature. Nevertheless, Cochrane (2011) contends that researchers may have overstepped by introducing an overwhelming multitude of factors, making it impractical and conceptually unwise to consider them collectively. He coins this phenomenon as a "factor zoo" and cautions against the indiscriminate use of numerous factors to explain the average cross-sectional returns.

This critique prompts a pertinent inquiry: which factors hold genuine significance? With such an abundance of potentially pertinent factors, this scenario presents itself as a high-dimensional conundrum. One pragmatic approach involves employing shrinkage techniques to enforce sparsity.

Feng et al. (2020) addressed the "factor zoo" challenge by employing a double-LASSO selection

procedure (see Belloni et al., 2014), favoring more parsimonious asset pricing models. Emphasizing out-of-sample predictability, Sun (2023) utilized the Ordered Weighted LASSO (OWL - Figueiredo and Nowak, 2016) to pinpoint factors capable of jointly explaining cross-sectional returns. Likewise, Freyberger et al. (2020) employed the Adaptive Group LASSO (see Huang et al., 2010) to "non-parametrically dissect the factor zoo".

These studies tackle the dimensionality issue of the factor zoo within a framework based on the Stochastic Discount Factor (SDF) model, deriving the shrinkage regression from the SDF error expression. Despite their ingenuity, this approach diverges from conventional methodologies for assessing the statistical significance of pricing factors in lower dimensions. Usually, classic methodologies involve spanning the proposed factor's returns against a given set of benchmark factors and scrutinizing the statistical significance of the intercept, as seen in studies by Jensen et al. (2023), Frazzini and Pedersen (2014), and Loh and Warachka (2012).

Similarly to approaches applied to the Stochastic Discount Factor (SDF) challenge, shrinkage regressions can be integrated into the spanning factors framework to address the dimensionality issue. One could use a shrinkage technique to factor spanning regressions and estimate the p-value of the intercept. Estimating a confidence interval for shrinkage regression coefficients isn't straightforward, given the bi-modal distributions of regressors' coefficients (Meinshausen and Bühlmann, 2006). However, refined techniques proposed by Meinshausen et al. (2009), build upon the foundation laid by Wasserman and Roeder (2009), offer promising avenues for estimating the statistical significance of the LASSO's intercept - with tailored adjustments for accommodating time-series data, one may emulate lower-dimensional solutions in high-dimensional environments.

The LASSO, known for shrinking coefficients toward zero, encounters a challenge in the presence of multicollinearity, prevalent in pricing anomalies' returns: distinguishing the genuine effects of individual regressors becomes difficult in such scenarios. Coefficients associated with correlated features may undergo shrinkage toward zero, even if one or both harbor true effects. Efforts by Freyberger et al. (2020) and Sun (2023) have deployed adaptations of the raw LASSO within the SDF framework to address this challenge. Another approach to mitigate major multicollinearity issues involves conducting a Variance Inflation Factor analysis before shrinkage regressions, similarly to Green et al. (2017).

The careful application of shrinkage is crucial when employing any of the aforementioned techniques, requiring researchers to calibrate the penalization parameter(s) carefully. Various methodologies have been advanced, each with distinct priorities; some prioritize predictability, exemplified by the K-fold Cross-Validation (CV) method (Stone, 1974), while others emphasize in-sample fit, as illustrated by the Bayesian Information Criterion (BIC) (Schwarz, 1978) and the Akaike Information

Criterion (AIC) (Akaike, 1974). It's worth noting, however, that none of these commonly employed criteria explicitly focus on ensuring a predefined level of sparsity.

Our contributions encompass three main fronts. Firstly, we introduce a framework for researchers dealing with the factor zoo issue using the classical factor-spanning method. This framework enables the distinction between the time series used to handle high multicollinearity between regressors, the one for estimating spanning regressions' intercepts to identify relevant factors, and the final one to predict out-of-sample returns. Secondly, we propose a new benchmark for evaluating factor selection models' out-of-sample predictability, inspired by the well-known Fama-French 3-factor model (Fama and French, 1993). This benchmark is stricter than just checking the Sharpe ratios of resulting hedge portfolios.

Additionally, we suggest methods to ensure a certain level of sparsity, both in shrinkage regressions and in assessing the relevance of pricing anomalies. In our study, we used LASSO regression for spanning regressions, employing traditional Bayesian Information and the proposed Fixed Regressors Criteria. Results slightly favored fixing 5 regressors as relevant.

The relevance of pricing factors was determined by examining the p-values of their spanning regression's intercept. When adhering to the conventional statistical significance approach, setting a significance level of 0.10 and utilizing factors with p-values below this threshold, results were underwhelming - particularly due to certain periods lacking any factors meeting this criterion. However, when we imposed a predefined level of sparsity, incorporating a fixed number of factors ranked by their p-values, we observed substantial enhancements in out-of-sample performance. We managed to build long-short portfolios with annualized Sharpe ratios - net of trading costs - reaching as high as 2.04.

Upon applying our proposed framework and sparsity-ensuring mechanisms to the factor-spanning challenge, we were able to observe how dissociating relevant periods impacted the factor zoo dimensionality problem. Traditionally, the literature has employed the same period, typically 120 months, for both factor selection and return prediction. However, our findings indicate that this dissociation resulted in more precise out-of-sample predictions, enhancing overall performance.

This paper is structured into five sections. Section 1 served as an introduction to the related literature and the achievements of this study. In section 2, we present the proposed framework, including a classical approach for accessing factor statistical significance and adjustments necessary for exploring its application in the high-dimensional factor zoo environment. Then, section 3 describes the data used in this paper, as well as the methodology for constructing the anomaly factors and test portfolios. In section 4, we present obtained results and propose a stricter benchmark for out-of-sample predictability. Finally, section 5 concludes the article, summarizing contributions and results.

2 Methodology

We evaluate the statistical significance of each factor individually through a series of spanning regressions, wherein we assess whether a combination of the other pricing anomalies can account for each factor's abnormal returns. This approach draws inspiration from studies by Jensen et al. (2023), Frazzini and Pedersen (2014), and Loh and Warachka (2012).

We delineate the classical approach for assessing the significance of pricing anomalies in section 2.1. Section 2.2 is dedicated to adapting this classical approach to the high-dimensional environment of the factor zoo. This includes addressing high multicollinearity (subsection 2.2.1), developing a methodology for estimating statistical significance (subsection 2.2.2), introducing two distinct criteria for setting the penalization parameter in shrinkage regressions (subsection 2.2.3), and proposing two criteria for assessing pricing factors' relevance within our framework (subsection 2.2.4). We elucidate the out-of-sample analysis conducted to verify the predictive capabilities of the selected factors in section 2.3 and introduce a rolling-window approach to the factor selection problem in section 2.4.

2.1 Assessing factors' significance in a low-dimensional environment

Denoting f as a proposed pricing anomaly, it is possible to assess f's significance by running a regression of its returns (ret$f$) over some benchmark factor model, as shown in Equation 1 below.

$$ret_f = \alpha + \sum_{j \in F_{bench}} \beta_j ret_{f_j}, \tag{1}$$

where ret_{f_j} represents the returns of a pricing factor that belongs to the set of relevant factors of the benchmark model (F_{bench}).

Take the classical Fama-French 3-factors model as an example benchmark model: ret_{f_j} would represent the returns of market, size, and value factors - see Fama and French (1993). Factor f would then be considered relevant in pricing returns if its alpha is relevant, i.e., if equation 1's intercept presents a low enough p-value. This idea is broadly used in the pricing anomalies literature, as in studies by Jensen et al. (2023), Frazzini and Pedersen (2014), and Loh and Warachka (2012).

2.2 Spanning factors' returns in a high-dimension environment

However, in recent years, the asset pricing literature has accepted as relevant a massive amount of factors - Hou et al. (2020) even compiled a data library of 447 published anomaly variables. Cochrane (2011) labeled this situation as a "zoo of factors", likening the vast variety of animals in a zoo to the myriad of factors available in the literature, each emitting a particular noise. Despite most of these

factors having shown some evidence of pricing impact, it is highly unlikely that all of them are jointly relevant in explaining the cross-section of returns.

We propose an approach that levels the playing field for every potentially relevant pricing anomaly, and while allowing for time-varying importance (see section 2.4), seeks to identify which factors have been the most statistically relevant. This is achieved through a slight modification of equation 1. The model is as follows:

$$ret_{f_{i,t}} = \alpha_{f_i} + \sum_{j \neq i, j \in F} \beta_j ret_{f_{j,t}} + \epsilon_{f_i} \tag{2}$$

where $ret_{f_{i,t}}$ is the return of factor f_i at time t, F is the set of all available pricing factors, α_i is the intercept of the regression, β_j is the linear coefficient for factor f_j, and ϵ_{f_i} is the error term.

The idea is to estimate each factor's alpha (α_{f_i}), calculate its p-value, and finally consider only the most statistically relevant. If done multiple times, using rolling windows of time series, this framework should capture how factor relevance has changed over time.

The factor zoo is a well-known high-dimensional environment, therefore it is not advisable to estimate equation 2 using a simple OLS regression. This is especially true as the object of interest will be the p-value of the intercept, and the more regressors we consider on the right-hand side of the model, the less likely it is to find a statistically significant intercept.

One possible solution is to estimate the model described in equation 2 using a shrinkage technique, such as the LASSO - see Tibshirani (1996).[1] However, two major concerns arise with the implementation of such a regression technique:

- Ordinarily, estimating the p-values for the LASSO's coefficients is a rather tricky exercise. We use a regression technique that will calculate an unbiased estimation of the LASSO intercept's p-value, accounting for situations where the data is presented as a time series;
- The LASSO regression is a technique sensitive to its penalization parameters. The literature provides some methodologies for properly setting it, like the information criteria and cross-validation, but we propose a way of ensuring the sparsity level desired.

2.2.1 Avoiding extreme multicollinearity

One of the primary concerns highlighted by Cochrane in his presidential address (Cochrane, 2011) is the extreme improbability that all proposed pricing factors are jointly relevant in pricing the cross-section of asset returns. This skepticism arises because some factors capture similar qualitative

[1] Results for the Elastic Net (Zou and Hastie, 2005) and Adaptive LASSO (Zou, 2006) will be available in the final version's Appendix.

information (e.g., illiquidity and zero trading days), some are combinations of others (e.g., current ratio and percentage change in current ratio), some vary only by a time frame of interest (e.g., 1, 6, 12, 36-month momentum), and some are even squared versions of others (e.g., beta and beta squared).

While extreme multicollinearity doesn't introduce bias in estimated slope coefficients, it does inflate their standard errors. For instance, severe multicollinearity poses a challenge for methods like the LASSO regression, which may arbitrarily drop one (or more) of the covariates from the model if they are highly correlated.

Green et al. (2017) proposes a classic solution to mitigate the effects of multicollinearity, suggesting the use of Variance Inflation Factors (VIFs). VIFs are defined as:

$$VIF_i = \frac{1}{1 - R_i^2}, \tag{3}$$

where R_i^2 is the unadjusted coefficient of determination for regressing the ith independent variable on the remaining ones.

The closer R_i^2 is to zero (one), the less (more) correlated the ith independent variable is to the others, implying that multicollinearity is less (more) likely to exist. The researcher then chooses a threshold for the higher VIF she will accept, disregarding all independent variables that exceed this set value - Green et al. (2017) removes factors for which $VIF_i > 7$. Used properly, this metric should capture how well the other factors' returns explain a given factor's returns.

We employ a procedure similar to Green et al. (2017) to address extreme factors' multicollinearity. However, we propose a few simple modifications to ensure that we do not introduce any look-ahead bias into the analysis. In their analysis, Green et al. (2017) calculate candidate factors' VIFs considering all available data. While this approach posed no major harm to their study focused on the cross-section of returns, our study allows for time-variant factors relevance. Removing factors based on full-sample multicollinearity introduces a clear source of look-ahead bias. Therefore, we conduct the VIF analysis in rolling windows, considering data available only before the time of interest.[2]

Furthermore, we adopt a method of removing high VIF factors one at a time, in decreasing order. The usual approach is to calculate the VIF for all independent variables at once, and then disregard the ones with high enough VIF. However, two variables could present a VIF higher than the threshold, and after removing one of them from the pool of independent variables, the other variable's VIF could potentially decrease to fit the acceptance level. To account for this behavior when addressing severe multicollinearity, we compute the VIF value for all candidate factors, remove only the one with the highest VIF, and then recalculate the VIF for all remaining factors - repeating the process until only factors with low enough Variance Inflation Factors survive.

[2]See section 2.4 for details on the time-variant framework.

2.2.2 Estimating intercept's p-value

In lower dimensions, where simple OLS regression can be used without significant concerns about overfitting, verifying the statistical significance of the intercept in equation 2 is straightforward. However, in higher dimensions, especially when applying shrinkage regressions, this task becomes more challenging.

Meinshausen et al. (2009) extend the concept of splitting the data into two parts, one for reducing the problem's dimensions and the other for applying classical variable selection techniques, originally proposed by Wasserman and Roeder (2009). They present a methodology for calculating the p-value for high-dimensional regressions. [3]

We follow the approach of Meinshausen et al. (2009). However, their methodology does not account for time-series data - therefore, our bootstrap procedure must consider the specificities of such data. To achieve this, we modify the simple bootstrap procedure used in the original study to employ a block bootstrap. For simplicity, we employ a non-overlapping blocks procedure (see Hall, 1985; Carlstein, 1986).

Taking B as the total number of bootstrap repetitions, for $b = 1, ..., B$:

- Randomly split the original dataset into two disjoint groups, D^b_{shrk} and D^b_{p-val}:
  - This split must be done in blocks due to the time-series nature of the data;
- Run the shrinkage regression, estimating the set of active predictors ($\tilde{F}^b$), using data from D^b_{shrk}:
  - Thereafter, $\tilde{F}^b = \{j, \beta j \neq 0\}$ after running the shrinkage in equation 2;
- Using only D^b_{p-val}, fit the selected factors in $\tilde{F}^b$ with Ordinary Least Squares and calculate the p-value for the intercept, $P^b_{\alpha_i}$.

This procedure leads to a total of B p-values $P^b_{\alpha_i}$ - and their suitable summary statistics are quantiles (see Meinshausen et al., 2009). For $\gamma \in (0, 1)$, define

$$\tilde{P}_{\alpha_i}(\gamma) = min\{1, q_\gamma([P^b_{\alpha_i}/\gamma; b = 1, ..., B])\}, \tag{4}$$

where $q_\gamma(\cdot)$ is the empirical γ-quantile function.

In equation 4, we provide a p-value for any fixed $0 < \gamma < 1$. However, selecting γ appropriately is not straightforward, and searching for its optimal value does not guarantee error control. To determine the final p-value, we can adopt an adaptive approach that selects a suitable quantile value using a data-driven methodology. Let $\gamma_{min} \in (0, 1)$, typically set to 0.05, be a lower bound for γ. We

[3] Wasserman and Roeder (2009) results are based on a single-split, whereas Meinshausen et al. (2009) proposed a multisplit method to avoid the randomness caused by data dependence.

then define the final p-value as:

$$P_{\alpha_i} = \min\{1, (1 - \log \gamma_{min}) \inf_{\gamma \in (0, \gamma_{min})} \tilde{P}_{\alpha_i}(\gamma)\} \tag{5}$$

We acknowledge that both the Bonferroni correction (Bonferroni, 1936) —applied after the bootstrap's final step — and the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) — employed after calculating the final p-values for all factors' intercepts — could be alternatives for mitigating data mining concerns. However, the results obtained in the empirical analysis, detailed in section 4, led us to conclude that these concerns were unnecessary. In general, lower p-values would have led to a potentially harmful decrease in out-of-sample portfolio performance.

2.2.3 Setting up the penalization parameter

The shrinkage penalty parameter merits special attention as a crucial component of the objective function, set by the researcher to promote sparsity and prevent overfitting. It dictates how severe the penalization will be, thus regulating the extent of shrinkage. Various techniques exist to aid in setting penalization parameters, including K-fold Cross-Validation (CV) and Information Criteria, such as Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC).

Cross-Validation, originally proposed by Stone (1974), partitions the data into training and validation sets, fits the model on the training set, and evaluates its performance on the validation set. It serves as a robust technique to circumvent overfitting and does not rely on assumptions about data distribution. Conversely, Bayesian Information Criterion (BIC) (Schwarz, 1978) and Akaike Information Criterion (AIC) (Akaike, 1974) weigh the trade-off between model fit and complexity, with BIC often favoring more parsimonious models than AIC.

One advantage of employing Information Criteria over Cross-Validation lies in their computational efficiency and ease of application to large datasets. However, Information Criteria hinge on strong data distribution assumptions and may be sensitive to their violation. In contrast, Cross-Validation offers robustness against such violations but demands computational resources and a substantial sample size for accurate prediction error estimates. As suggested by Freyberger et al. (2020), we posit that, within customary methodologies, BIC emerges as more suitable for our application.

In the process of traversing the factor zoo across different periods, there is no guarantee that the number of factors surviving the shrinkage process in spanning regressions will remain constant. This is particularly significant given that our methodology relies on assessing intercepts' p-values: the greater the number of factors with non-zero β_j coefficients in equation 2, the less likely it is for α_{f_i} to differ significantly from zero.

To address this challenge, we propose a distinct criterion for setting the penalization parame-

ters in shrinkage regressions. The objective is to ensure that a predetermined number of regressors (as determined by the researcher) will possess non-zero coefficients. This criterion proves especially advantageous in the realm of factor-based asset-pricing models, offering *ad-hoc* means to ensure the implicit sparsity assumption inherent in models like the Fama-French 3/5, Carhart, and q-4 factors (Fama and French, 1993, 2015; Carhart, 1997; Hou et al., 2015), by selecting a penalty parameter that precisely returns the desired number of non-zero coefficient regressors.

The Fixed Regressors Criterion (FRC) determines the penalty parameter (λ) through the following algorithm:

- Initialize an array of candidate values and perform the shrinkage regression on them;
- Ensure that the desired number of final regressors is included in the range of candidate λ's:
  - If the number of selected factors for the highest (lowest) candidate λ is too low (high), adjust for lower (higher) values. Re-run the regression for this new set of candidate λ's;
- If no value precisely returns the desired number of regressors, narrow the range:
  - Set the new highest (lowest) possible value as the higher (lower) penalization that yields fewer (more) than the desired number of factors. Re-run the regression for this new array of candidate penalization parameters;
- Repeat the above steps until at least one candidate value returns exactly the desired number of factors with non-zero coefficients:
  - As multiple penalization values may return the set number of relevant regressors, select the median value as the penalization parameter.

While implementing the FRC, the researcher should consider the specifics of the chosen estimation technique. For instance, the LASSO does not guarantee a monotonically increasing number of selected factors as the penalization parameter decreases. Additionally, although highly unlikely, there is no assurance that the exact number of desired regressors will be chosen for any set value of the penalty parameter. Given the numerous regressions involved, it's probable that an isolated unlikely situation will occur, and the researcher should be prepared to address it. We recommend selecting the median value from the highest penalization sequence of candidate values to address the non-monotonicity issue and ceasing the search for the perfect lambda after several iterations, opting for the highest penalization that yields one more relevant factor.[4]

[4]Regarding the extra factor, it is up to the researcher to decide what to do. If the data is scaled and the absolute values of the coefficients are comparable, we suggest disregarding the factor with the lower $|\beta|$, as suggested by Sun (2023). However, disregarding selected regressors means not necessarily being supported by the chosen regression properties.

2.2.4 Accessing factor relevance

The estimation of factors' intercepts p-values emphasizes the relevance of factors with smaller p-values. A statistically correct approach for selecting relevant factors involves stipulating a significance level and considering factors with $P_{\alpha_{f_i}}$ lower than that level.

However, this approach may not be optimal in environments where predictability is challenging, resulting in relatively large p-values. This is particularly true in fields such as return predictability, where regressions often have less-than-ideal explanatory power. In some cases, none of the candidate factors may generate p-values small enough to be considered significant. Conversely, there may even be instances where too many factors survive the shrinkage process, leading to an unwieldy number of factors to account for.

Given these concerns, we propose an alternative approach to assess factor relevance in pricing assets' returns. Rather than fixing a significance level, we fix the number of factors considered relevant, allowing the researcher the flexibility to set the desired number of factors in their model.

With this approach, we can maintain the silent sparsity assumption of classic asset pricing models, where only a handful of factors are considered. However, instead of selecting factors solely based on financial or economic intuition, we rely on the data, using statistical significance to identify relevant factors. Testing for more extreme sparsity levels is also possible, such as considering only the single factor with the lowest p-value.

2.3 Out-of-sample analysis

We assess the predictability of the chosen factors through an out-of-sample analysis, inspired by the methodology outlined in Freyberger et al. (2020).

We use the returns of hedge portfolios constructed from the predictions of test assets' returns based on the selected factors. Freyberger et al. (2020) employ the returns of the selected factors, delayed by one period, as predictors for test asset returns in simple OLS regressions conducted over rolling windows of 120 months. Subsequently, the OLS coefficients are utilized to forecast the returns of test assets one period ahead. A trading strategy is then formulated, involving hedge portfolios: assets in the top decile of predicted returns are bought, while those in the bottom decile are sold. If the strategy yields significant alpha, the variable selection is deemed successful.

This kind of one-period delay approach is common in the literature and tacitly assumes factor momentum, as it implies that the best approximation for the factor return at time t is the return at $t-1$. Given the widespread acceptance of factor momentum (see Houweling and Van Zundert, 2017; Gupta and Kelly, 2019), we find this assumption reasonable, particularly considering the interpretability of the results.

We adopt the same methodology as Freyberger et al. (2020) but allow for variations in the rolling window sizes. The rationale behind this choice is that different window lengths may carry distinct economic implications. A shorter window, where observations are on average closer to the return to be predicted, may offer less overall information but closely reflect market behavior around the estimation period. Conversely, longer time series will capture the relationship between selected factors and test assets over an extended horizon, albeit less "instantly" related to the prediction moment, yet benefiting from greater statistical power due to the larger dataset.

2.4 Time-varying factor selection framework

Our framework empowers researchers with creative freedom, enabling them to consider various distinct windows of interest. This approach may support diverse financial intuitions while meticulously avoiding look-ahead biases.

Drawing from the methodology proposed by Freyberger et al. (2020), we advocate for the separation of rolling windows into three distinct phases: one for addressing multicollinearity, another for estimating relevant factors, and a final one for conducting out-of-sample prediction regressions - denoted as RW_{VIF}, RW_{shrk}, and RW_{pred}, respectively.

This segmentation facilitates a comprehensive exploration of the factor zoo in a time-varying manner. Specifically, it allows for the investigation of different time frames to select relevant factors and understand the relationship between the returns of these factors and the returns of test assets.

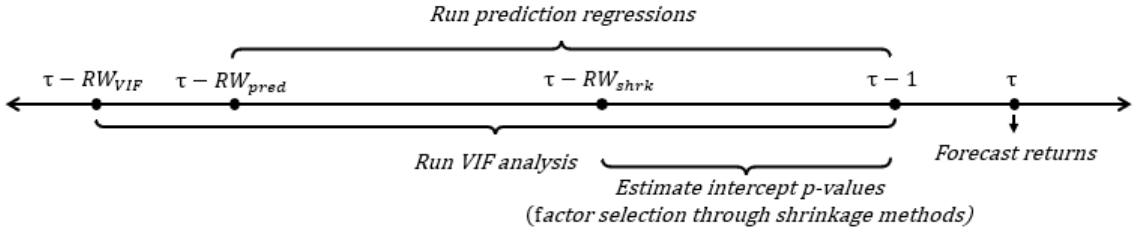

Figure 1: Variable section rolling-window framework scheme

Illustration of the proposed time-variant factor selection framework. In this scenario, for forecasting test assets' returns for time τ, VIFs are computed with data from $\tau - 1$ to $\tau - RW_{VIF}$, factor selection considers data from $\tau - 1$ to $\tau - RW_{shrk}$, and assets' returns at time t are predicted using a time series from $\tau - 1$ to $\tau - RW_{pred}$.

The schematic representation of the proposed variable selection framework is presented in figure 1, and goes as follows:

- To forecast test assets' returns at time τ, run the VIF procedure (see subsection 2.2.1) using data from $\tau - 1$ to $\tau - RW_{VIF}$;
- For every surviving factor f_i, estimate $\alpha_{f_i,\tau-1}$ (see equation 2) using data from $\tau - 1$ to $\tau - RW_{shrk}$, find its corresponding p-value (following subsection 2.2.2), and select the relevant

factors - criteria presented in subsection 2.2.4;

- Over a different time series, from $\tau - 1$ to $\tau - RW_{pred}$, regress the returns of selected factors against every test asset's returns, delayed by one period, as described in section 2.3;
- Finally, use the OLS coefficients to project test assets' returns for time τ:
  - Build neutral long-short portfolios based on predicted returns, buying (selling) the top (bottom) decile;
- Repeat the process for all possible time periods, storing the long-short portfolio returns;
- Calculate a (some) metric(s) for validating the strategy's performance - like the Sharpe Ratio - see Sharpe (1998).

This methodology facilitates the examination of the spanning factors problem by ensuring a clear separation between the time series used for major multicollinearity treatment, the one to compute p-values and select relevant factors, and the time series utilized to forecast returns based on that selection. It offers versatility, accommodating studies that focus on stable factor models spanning decades of data, as well as those focusing on shorter periods such as intraday estimations.

3 Data

We utilize data from the Center for Research in Security Prices (CRSP) and Compustat databases, covering the period from January 1980 to December 2021. The dataset comprises 504 months of data on all common stocks listed on NYSE, AMEX, and NASDAQ, encompassing the same 80 characteristics used by Sun (2023). Risk-free rate and market excess returns are obtained from Kenneth French's online data library. In section 3.1, we detail how the anomaly factors were calculated and provide an overview of the zoo of factors, while in section 3.2, we explain how we constructed the test assets.

3.1 Anomaly factors

Factors considered in this study are constructed based on published asset pricing anomalies, which are defined by Brennan and Xia (2001) as *"statistically significant differences between the realized average returns associated with certain characteristics of securities, or on portfolios of securities formed based on those characteristics, and the returns that are predicted by a particular asset pricing model"*.

In addition to the market factor, we examine 80 additional characteristics as possible regressors - see table 1. We exclude micro stocks with a market capitalization smaller than the 20th percentile of NYSE-listed stocks.[5]

[5]Micro stocks are classified monthly.

Table 1: Anomaly factors

This table lists all used factors. The abbreviation is consistent with Green et al. (2017) and Sun (2023). Detailed information is available at Green et al. (2017).

Abbreviation	Description	Abbreviation	Description
absacc	Absolute accruals	mom1m	1-month momentum
acc	Working capital accruals	mom36m	36-month momentum
aeavol	Abnormal earnings announcement volume	mom6m	6-month momentum
agr	Asset growth	ms	Financial statement score
baspread	Bid-ask spread	mve	Size
beta	Beta	mve_ia	Industry adjusted size
betasq	Beta squared	nincr	Number of earnings increases
bm	Book-to-market	operprof	Operating profitability
bm_ia	Industry adjusted book-to-market	pchcapx_ia	Industry adjusted % change in capital expenditures
cash	Cash holding	pchcurrat	% change in current ratio
cashdebt	Cash flow to debt	pchdepr	% change in depreciation
cashpr	Cash productivity	pchgm_pchsale	% change in gross margin - % change in sales
cfp	Cash flow to price ratio	pchquick	% change in quick ratio
cfp_ia	Industry adjusted cfp	pchsale_pchinvt	% change in sale - % change in inventory
chatoia	Industry adjusted change in asset turnover	pchsale_pchrect	% change in sale - % change in A/R
chcsho	Change in share outstanding	pchsale_pchxsga	% change in sale - % change in SG&A
chempia	Industry adjusted change in employees	pchsaleinv	% change in sales-to-inventory
chinv	Change in inventory	pctacc	Percent accruals
chmom	Change in 6-month momentum	pricedelay	Price delay
chpmia	Industry adjusted change in profit margin	ps	Financial statement score
chtx	Change in tax expense	quick	Quick ratio
cinvest	Corporate investment	retvol	Return volatility
currat	Current ratio	roaq	Return on assets
depr	Depreciation	roavol	Earning volatility
dolvol	Dollar trading volume	roeq	Return on equity
dy	Dividend-to-price	roic	Return on invested capital
ear	Earnings announcement return	rsup	Revenue surprise
egr	Growth in common shareholder equity	salecash	Sales to cash
ep	Earnings-to-price	saleinv	Sales to inventory
gma	Gross proditability	salerec	Sales to receivables
grcapx	Growth in capital expenditure	sgr	Sales growth
grltnoa	Growth in long term net operating assets	sp	Sales-to-price
hire	Employee growth rate	std_dolvol	Volatility of liquidity (dollar trading volume)
idiovol	Idiosyncratic return volatility	std_turn	Volatility of liquidity (share turnover)
ill	Illiquidity	stdacc	Accrual volatility
invest	Capital expenditure and inventory	stdcf	Cash flow volatility
lev	Leverage	tang	Debt capacity/firm tangibility
lgr	Growth in long term debt	tb	Tax income to book income
maxret	Max daily return	turn	Share turnover
mom12m	12-month momentum	zerotrade	Zero trading days

We compute the factors as the spread returns between top and bottom decile portfolios, controlling for size. This approach is akin to a more tail-oriented version of Fama and French (1993)'s methodology, which uses deciles instead of 30% percentiles. For our estimation, the key consideration is whether the factor survives the dimensionality-lowering procedure, and all factors are calculated on a high-minus-low basis - regardless of whether they are characterized as low-minus-high. We demean and adjust all factors to share the same standard deviation as the market factor. This facilitates interpretation and makes the magnitudes of estimated coefficients comparable. Finally, we remove characteristics that cannot produce factors for all available dates.

3.2 Test assets

The literature offers different perspectives on the ideal set of test assets for asset pricing models. Some scholars advocate using individual stocks, while others promote the utilization of sorted portfolios. While individual stocks have been used in some studies (see Harvey and Liu (2021) and Lewellen (2015)), Feng et al. (2020) have argued that characteristic-sorted portfolios have more stable betas, generally present better signal-to-noise ratios, and are more protected from missing data issues.[6] In line with these arguments, we generate our test assets by sorting the stocks into portfolios based on their factors' characteristics.

Following the methodology of Sun (2023), we construct a comprehensive set of bivariate sorted portfolios as our test assets, in line with the approach proposed by Feng et al. (2020) and Freyberger et al. (2020). The approach involves creating all 5×5 bivariate sorted portfolios, formed by intersecting stocks' size with each of the 80 characteristics considered in the previous subsection. The construction process is similar to that described in section 3.1 for anomaly factors, but resulting portfolios are both long-only and less extreme.

At the end of the process, any bivariate portfolio that fails to generate diversified portfolios for all dates of interest will be excluded from the set of test assets. Therefore, we end up using a total of 1896 diversified portfolios as the full set of test assets.

4 Empirical analysis

In this section, we apply the methodology outlined in section 2 to scrutinize the factor zoo, aiming to reduce its dimensionality by examining the interplay of factor returns. Each factor's returns are compared against all others using the LASSO methodology to identify factors capable of generating statistically significant intercepts in the spanning regression (refer to equation 2). We set the rolling window for the VIF methodology (RW_{VIF}) at 240 months.[7] We also explore three different window lengths (120, 180, and 240 months) for both RW_{shrk} and RW_{pred}.

To determine the shrinkage penalization parameter, we consider 100 possible values, with initial values set at 10^{-4} and 10^{-1} for the lowest and highest possible penalization parameters, respectively. The block bootstrap procedure is employed with eighty repetitions of blocks of five observations.

We report results for two approaches for selecting the shrinkage penalization parameter: the classic Bayesian Information Criterion (BIC) and our proposed Fixed Regressors Criterion (FRC),

[6] Fama and French (2008) and Hou et al. (2015) have also advocated using of sorted portfolios as test assets.

[7] Threshold for accepted VIF value is 10 - a little more permissive than Green et al. (2017).

which fixes five regressors in the spanning regression.[8]

We commence by discussing the results of the VIF multicollinearity treatment in section 4.1, followed by addressing the high-dimensionality issue in section 4.2. Subsequently, we present the out-of-sample performance of our framework in section 4.3. Finally, section 4.4 proposes a new benchmark for assessing the out-of-sample performance of the resulting portfolios.

4.1 Accounting for multicollinearity

The empirical analysis begins by addressing potential issues arising from extreme multicollinearity among pricing anomalies. We address this concern by applying the Variance Inflation Factor (VIF) criterion, which evaluates the R^2 of regressions where all other factors' returns explain each factor's returns (refer to subsection 2.2.1).

The results of this procedure, conducted with a rolling window of 240 periods,[9] are presented in this section. Figure 2 illustrates the number of factors that survive this process over time, indicating a relatively constant count of accepted factors throughout the sample period. The accepted factors range from 60 to 64 for most of the studied period, with an average elimination of 18.62 factors. Notably, there is a discernible increase in the number of factors accepted by the VIF analysis after the second half of 2019, suggesting a potential decrease in overall multicollinearity among pricing anomalies.

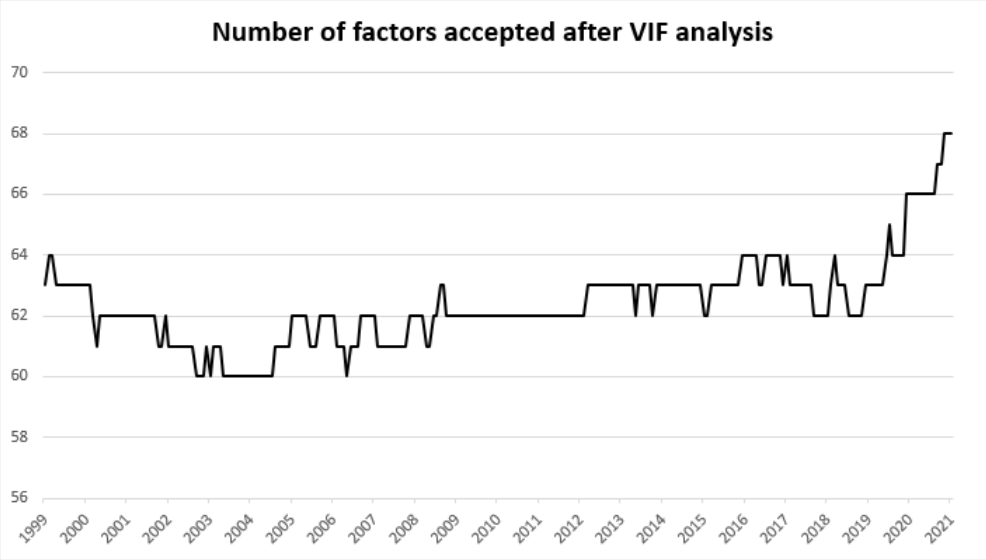

Figure 2: Pricing anomalies accepted by the Variance Inflation Factor control trough time
The plot reports the number of factors that survived the VIF analysis throughout all the available periods.

Complementing the analysis, table 2 lists all the anomalies rejected after the VIF procedure for

[8] Although our results do not account for trading and slippage costs, average monthly turnover is reported.

[9] In the final version, results using distinct periods will be presented as robustness checks.

at least one period. It reveals that seven candidate pricing factors were consistently disregarded due to multicollinearity, while seventeen factors were rejected in at least 80% of the time series.

Furthermore, our results demonstrate notable stability in identifying factors prone to multicollinearity issues. Nineteen out of the thirty-one pricing anomalies that encountered problems at least once failed to survive the VIF procedure more than 60% of the time.[10][11]

Table 2: Summary of pricing anomalies rejected by the Variance Inflation Factor control

This table lists all factors rejected by the VIF analysis at least once, and reports the number of times the factors presented multicollinearity issues - and the percentage of available periods they were rejected. The abbreviation is consistent with Green et al. (2017) and Sun (2023).

Factor	Rejections	Percentage
baspread	265	100.0%
beta	265	100.0%
betasq	265	100.0%
cash	265	100.0%
lev	265	100.0%
retvol	265	100.0%
zerotrade	265	100.0%
idiovol	262	98.9%
mom6m	262	98.9%
dy	260	98.1%
stdcf	258	97.4%
std_turn	253	95.5%
turn	251	94.7%
salecash	245	92.5%
ill	242	91.3%
sp	232	87.5%

Factor	Rejections	Percentage
quick	219	82.6%
roeq	176	66.4%
agr	160	60.4%
stdacc	125	47.2%
pchsaleinv	95	35.8%
roaq	91	34.3%
ep	72	27.2%
currat	46	17.4%
bm	30	11.3%
gma	24	9.1%
roavol	21	7.9%
absacc	10	3.8%
pchsale_pchinvt	6	2.3%
maxret	4	1.5%
invest	1	0.4%

4.2 Reducing dimensionality

After addressing high multicollinearity, our next step was to select relevant factors and reduce the dimensionality of the factor zoo. As outlined in subsection 2.2.2, we estimated the p-value of the intercept in spanning regressions - see equation 2 - to guide our selection process.

In our empirical analysis, we employed the LASSO as the shrinkage estimator and considered both methodologies detailed in subsection 2.2.3 for setting its penalty parameter: the classic BIC and the proposed FRC, fixing five regressors (labeled f5). P-values were estimated using three distinct rolling windows (RW_{shrk}) of 120, 180, and 240 months.

Table 3 summarizes the results obtained using our proposed methodology, shedding light on several interesting aspects. Firstly, our methodology is notably conservative in estimating p-values, with only a small average number of factors yielding valid p-values, reaching approximately 13.4% of

[10] Exploring more the reasons behind factors multicollinearity is a possible aspect to develop. It is possible to extract which factors were more relevant in the VIF analysis to find out which variables were more relevant in explaining each removed factor.

[11] Reporting a table with the average VIF for each factor may be a nice addition. We could point out which factors are less explained by the others.

candidates. Moreover, the number of factors with valid p-values appears to decrease as the length of the time series used for regression increases.[12]

Table 3: Spanning factors returns shrinkage outcome

This table summarizes the outcomes of applying the LASSO regressions over the factors spanning regressions, reporting the average number of valid p-values, i.e., p-values under 1.0, the average number of p-values under the threshold of 0.10, and the percentage of time that any factor presents associated p-value smaller than 0.10, setting shrinkage penalization parameter through BIC of FRC (fixing 5 regressors - f5), and different values for RW_{shrk}.

Penalty style	RW_{shrk}	Average # of valid p-values	Average # of p-values < 0.10	% of periods with zero p-values < 0.10
BIC	120	10.00	1.09	34%
BIC	180	9.37	1.18	32%
BIC	240	7.15	1.24	34%
f5	120	10.69	1.24	30%
f5	180	9.99	1.20	35%
f5	240	7.15	1.12	36%

Secondly, the average number of factors with relatively low p-values, under 0.10, is small across all combinations of penalty styles and rolling windows analyzed. This trend may be attributed to the third insight gleaned from table 3: a significant percentage of periods (at least 30% of the time) saw no factor exhibiting an intercept p-value below 0.10.

Considering all the aforementioned aspects, employing the fixed 5 regressors criterion for setting the shrinkage penalization and conducting regressions using 120-period windows yielded the most favorable results. This approach yielded the highest average number of valid p-values, the highest average of factors with low p-values, and the lowest percentage of periods with no factor exhibiting a p-value below 0.10. However, there is no clear indication that a particular penalty style or RW_{shrk} is optimal, as there is no evidence of a relationship between RW_{shrk} and the overall level of statistical significance of the factors' p-values. Additionally, there is no discernible time-variant behavior, as illustrated by plots at the appendix.

An elevated number of periods without statistically significant pricing factors could significantly undermine the performance of a hedge portfolio constructed based on the forecasted returns of test assets — a common approach in the literature applied in this study— as it would frequently lead to deallocation. Moreover, the relatively low statistical significance is not entirely unexpected, given that time-varying asset pricing operates within a particularly noisy environment. Therefore, a method that rejects all pricing factors more than 30% of the time may be overly selective.

Sailing through the turbulent sea of pricing factors, it might be intriguing to focus on the most promising options, regardless of their individual p-values. As proposed in subsection 2.2.4, rather

[12] Meinshausen et al. (2009) methodology provides a conservative approach to family-wise error rate (FWER) control, similar in spirit to Holm (1979). If many null-hypotheses rejections were to happen, the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) could be considered.

than setting a threshold value for the factor p-value, we can instead designate a desired number of factors to be considered relevant and select those with the highest statistical significance, i.e., the lowest p-values — even if they exceed the typical acceptance threshold. This proposal, involving an *ad-hoc* determination of the number of relevant factors, also aligns with the silent sparsity assumption of classic asset pricing models. Just as the literature accepts classic models like Fama-French (see Fama and French, 1993, 2015), or Carhart and q-4 factors (see Carhart, 1997; Hou et al., 2015), with only a handful of pricing factors, why not let a statistical model dictate comparable sparsity levels without relying on prior qualitative information?[13]

4.3 Out-of-sample results

In this subsection, we will present the results obtained by applying the methodology explained in section 2.3. We will begin by presenting the results obtained using the BIC criterion for setting the penalization parameter in the spanning regression (subsection 4.3.1). Subsequently, we will discuss the results obtained by fixing five regressors according to the proposed FRC (see subsection 4.3.3). In each exposition, we will initially present the results obtained using only factors whose p-values are lower than 0.10, followed by the presenting results using one, three, or five factors with the lowest p-value(s).

4.3.1 Penalizing through BIC

Results obtained using the BIC for penalization in the spanning regressions, with a significance threshold set at p-values under 0.10, are presented in table 4, showcasing not that impressive performances.

At first glance, an uninformed reader might find the overall gross Sharpe Ratios above unity somewhat impressive. However, a closer examination reveals that these results indicate, at most, a promising direction rather than significant performance gains. Moreover, they fail to outperform the results obtained with a more suitable benchmark, as later proposed in section 4.4.

The relatively underwhelming performance can be attributed to the method for selecting relevant factors, a p-value threshold, which results in numerous periods where constructing long-short portfolios becomes infeasible due to the absence of relevant factors for predicting test asset returns.[14]

[13] We intend to explore more the individual statistical significance of each factor. In this version, we focused on the predictability results, however, in upcoming development, we will present information on each factor's average p-values, how statistical significance varied over time, how the *nature* of high-significant factors change over time, and other peer suggestions.

[14] As demonstrated in table 3, when employing the BIC penalization, there are instances where no factor exhibits a p-value under 0.10, accounting for at least 32% of the time.

Table 4: Out-of-sample results - BIC penalty and p-values under 0.10

This table reports out-of-sample monthly mean returns and associated standard deviations, annualized Sharpe ratios, and monthly average turnover of hedge portfolios going long-short the 10% of stocks with the highest-lowest predicted returns, considering factors with associated p-values lower than 0.10, setting shrinkage penalization parameter through BIC, for different combinations of RW_{shrk} and RW_{pred}.

RW_{shrk}	120			
RW_{pred}	Mean	SD	Sharpe	Turnover
240	0.0088	0.0205	1.49	1.38
180	0.0085	0.0223	1.32	1.38
120	0.0088	0.0237	1.29	1.41

RW_{shrk}	180			
RW_{pred}	Mean	SD	Sharpe	Turnover
240	0.0086	0.0246	1.21	1.39
180	0.0092	0.0258	1.23	1.40
120	0.0096	0.0261	1.27	1.40

RW_{shrk}	240			
RW_{pred}	Mean	SD	Sharpe	Turnover
240	0.0100	0.0247	1.40	1.35
180	0.0104	0.0254	1.41	1.36
120	0.0102	0.0262	1.35	1.34

Consequently, there are many instances where the returns of the hedge portfolios are zero, thus diluting the mean returns and undermining the obtained Sharpe ratios.[15]

Table 5 presents the results obtained when selecting a predetermined number of factors as relevant, using the lowest p-values obtained for each period for ranking, revealing more promising out-of-sample performance.

Hedge portfolios' mean returns reported in table 5 exhibit more significant values, resulting in improved Sharpe ratios, particularly when considering a smaller number of regressors, as shown in the first two columns. Generally, smaller values for RW_{shrk} and higher values for RW_{pred} tend to perform better. However, an intriguing aspect derived from the presented results is that there is no clear optimal value for RW_{shrk} and RW_{pred} to achieve the highest Sharpe ratios. The best value (1.95) is found in the pair ($RW_{shrk} = 120$; $RW_{pred} = 240$), while the third-best result (1.85) is observed in the combination of ($RW_{shrk} = 240$; $RW_{pred} = 180$).

Regarding the number of considered factors, our results suggest that researchers should aim for a high level of sparsity, selecting at most three factors when pricing asset returns. Most remarkably, the best outcomes are achieved when the maximum degree of sparsity is enforced. This represents a robust and unexpected finding, indicating that within our environment and proposed framework, the

[15]Since the hedge portfolios are constructed with 100% long-short positions, assigning a zero return when no relevant factor is selected is the appropriate approach. Additionally, attributing a risk-free rate return when no factor is chosen is not feasible in our Sharpe ratio analysis, as all long-short portfolios have cash available for investment at all times due to their construction methodology.

Table 5: Out-of-sample results - BIC penalty and 1, 3, and 5 lowest p-values

This table reports out-of-sample monthly mean returns and associated standard deviations, annualized Sharpe ratios, and monthly average turnover of hedge portfolios going long-short the 10% of stocks with the highest-lowest predicted returns, considering factors with the lowest 1, 3, and 5 associated p-values, setting shrinkage penalization parameter through BIC, for different combinations of RW_{shrk} and RW_{pred}.

Lowest p-value

RW_{shrk}	120			
RW_{pred}	Mean	SD	Sharpe	Turnover
240	0.0130	0.0231	1.95	1.40
180	0.0135	0.0247	1.89	1.41
120	0.0126	0.0254	1.72	1.39

RW_{shrk}	180			
RW_{pred}	Mean	SD	Sharpe	Turnover
240	0.0121	0.0235	1.79	1.38
180	0.0129	0.0252	1.78	1.36
120	0.0130	0.0258	1.74	1.37

RW_{shrk}	240			
RW_{pred}	Mean	SD	Sharpe	Turnover
240	0.0119	0.0252	1.64	1.37
180	0.0121	0.0268	1.56	1.41
120	0.0120	0.0276	1.51	1.48

3 lowest p-values

RW_{shrk}	120			
RW_{pred}	Mean	SD	Sharpe	Turnover
240	0.0111	0.0231	1.67	2.15
180	0.0113	0.0245	1.61	2.20
120	0.0116	0.0260	1.54	2.30

RW_{shrk}	180			
RW_{pred}	Mean	SD	Sharpe	Turnover
240	0.0122	0.0261	1.63	2.10
180	0.0129	0.0274	1.62	2.06
120	0.0132	0.0279	1.64	2.12

RW_{shrk}	240			
RW_{pred}	Mean	SD	Sharpe	Turnover
240	0.0140	0.0265	1.83	2.17
180	0.0144	0.0271	1.85	2.17
120	0.0144	0.0277	1.80	2.27

5 lowest p-values

RW_{shrk}	120			
RW_{pred}	Mean	SD	Sharpe	Turnover
240	0.0108	0.0271	1.38	2.52
180	0.0108	0.0274	1.37	2.54
120	0.0111	0.0281	1.37	2.64

RW_{shrk}	180			
RW_{pred}	Mean	SD	Sharpe	Turnover
240	0.0115	0.0283	1.41	2.45
180	0.0118	0.0292	1.40	2.45
120	0.0119	0.0302	1.36	2.54

RW_{shrk}	240			
RW_{pred}	Mean	SD	Sharpe	Turnover
240	0.0115	0.0297	1.34	2.49
180	0.0111	0.0299	1.29	2.50
120	0.0114	0.0301	1.31	2.55

noise inherent in the factor zoo is substantial enough to have better results potentially being yielded when considering only the most relevant pricing factor over the shrinkage period.

Finally, it is noteworthy that turnover also increases with the number of factors considered. This is a consequence of the persistence of the selected factor(s). Despite allowing for monthly changes, this framework for factor selection does not impose frequent and significant alterations in the relevant factors. This observation supports the notion of opting for a more sparse factor asset-pricing model, as lower turnovers translate to more favorable trading conditions and reduced slippage costs.

4.3.2 Using FRC for 5 factors

Similar to its BIC counterpart, fixing five regressors in the spanning regressions before selecting factors with p-values under 0.10 also yields unimpressive results, as shown in table 6.

Comparing table 6 to table 4, the overall results are very similar, with no clear indication of which combination of RW_{shrk} and RW_{pred} is optimal.

Table 7 presents the out-of-sample results obtained by combining the fixed five regressors criterion for the spanning regressions and selecting the lowest one, three, and five p-values for relevant factors. Once again, the results are similar to those in table 5. However, using the f5 criterion, we were able to achieve portfolios with Sharpe ratios greater than 2.0, surpassing the best out-of-sample Sharpe ratio obtained in subsection 4.3.1: 1.95.

Table 6: Out-of-sample results - FRC (f5) penalty and p-values under 0.10

This table reports out-of-sample monthly mean returns and associated standard deviations, annualized Sharpe ratios, and monthly average turnover of hedge portfolios going long-short the 10% of stocks with the highest-lowest predicted returns, considering factors with associated p-values lower than 0.10, setting shrinkage penalization parameter through FRC, fixing 5 regressors, for different combinations of RW_{shrk} and RW_{pred}.

RW_{shrk}	120			
RW_{pred}	Mean	SD	Sharpe	Turnover
240	0.0090	0.0219	1.42	1.41
180	0.0089	0.0229	1.35	1.40
120	0.0084	0.0241	1.21	1.44

RW_{shrk}	180			
RW_{pred}	Mean	SD	Sharpe	Turnover
240	0.0079	0.0246	1.11	1.42
180	0.0080	0.0261	1.06	1.44
120	0.0087	0.0268	1.12	1.45

RW_{shrk}	240			
RW_{pred}	Mean	SD	Sharpe	Turnover
240	0.0098	0.0236	1.44	1.29
180	0.0099	0.0245	1.41	1.32
120	0.0100	0.0256	1.35	1.33

Again, the results advocate for a high sparsity level, as choosing only one factor at a time appears to be the superior approach — especially considering the lower turnover imposed. Overall, results tend to improve when considering smaller shrinkage windows ($RW_{shrk} \in [120, 180]$), longer prediction time series ($RW_{pred} \in [240, 180]$), and a lower number of factors considered (only the lowest p-value). Nevertheless, this is not always the case: good results were obtained when using $RW_{shrk} = 240$ with the three lowest p-values —although those results impose a significant increase in average turnover.

A last notable observation is that the results appear to be relatively insensitive to the choice of the rolling window parameters: the sparsity level imposed by the researcher for factor selection seems to have greater importance. Our findings suggest that, in a noisy environment, such as the zoo of factors, *sparsity matters*.

4.4 Fair comparison

Practitioners often use an annualized Sharpe ratio, gross of trading costs, above one as a "rule of thumb" for considering hedge portfolio returns interesting. However:

- No "rule of thumb" should be accepted in rigorous scientific research, and;
- Given that all the factor anomalies considered in our study were reported as statistically significant return predictors in other studies, abnormal results may be a byproduct of their documented

Table 7: Out-of-sample results - FRC (f5) penalty and 1, 3, and 5 lowest p-values

This table reports out-of-sample monthly mean returns and associated standard deviations, annualized Sharpe ratios, and monthly average turnover of hedge portfolios going long-short the 10% of stocks with the highest-lowest predicted returns, considering factors with the lowest 1, 3, and 5 associated p-values, setting shrinkage penalization parameter through FRC, fixing 5 regressors, for different combinations of RW_{shrk} and RW_{pred}.

Lowest p-value

RW_{shrk}	120			
RW_{pred}	Mean	SD	Sharpe	Turnover
240	0.0131	0.0226	2.01	1.38
180	0.0131	0.0241	1.88	1.38
120	0.0123	0.0248	1.72	1.35

RW_{shrk}	180			
RW_{pred}	Mean	SD	Sharpe	Turnover
240	0.0135	0.0229	2.04	1.36
180	0.0142	0.0243	2.03	1.34
120	0.0144	0.0250	2.00	1.36

RW_{shrk}	240			
RW_{pred}	Mean	SD	Sharpe	Turnover
240	0.0121	0.0246	1.71	1.40
180	0.0123	0.0262	1.62	1.43
120	0.0120	0.0270	1.54	1.50

3 lowest p-values

RW_{shrk}	120			
RW_{pred}	Mean	SD	Sharpe	Turnover
240	0.0117	0.0232	1.75	2.15
180	0.0113	0.0244	1.60	2.18
120	0.0117	0.0257	1.57	2.30

RW_{shrk}	180			
RW_{pred}	Mean	SD	Sharpe	Turnover
240	0.0120	0.0251	1.65	2.23
180	0.0122	0.0265	1.60	2.20
120	0.0125	0.0277	1.56	2.28

RW_{shrk}	240			
RW_{pred}	Mean	SD	Sharpe	Turnover
240	0.0143	0.0259	1.91	2.18
180	0.0143	0.0266	1.86	2.17
120	0.0141	0.0273	1.79	2.25

5 lowest p-values

RW_{shrk}	120			
RW_{pred}	Mean	SD	Sharpe	Turnover
240	0.0106	0.0256	1.44	2.51
180	0.0108	0.0260	1.45	2.54
120	0.0108	0.0273	1.37	2.63

RW_{shrk}	180			
RW_{pred}	Mean	SD	Sharpe	Turnover
240	0.0104	0.0271	1.33	2.48
180	0.0107	0.0286	1.29	2.50
120	0.0108	0.0301	1.25	2.60

RW_{shrk}	240			
RW_{pred}	Mean	SD	Sharpe	Turnover
240	0.0117	0.0288	1.41	2.48
180	0.0110	0.0297	1.28	2.46
120	0.0112	0.0305	1.27	2.52

relevance.

Merely observing the resulting Sharpe ratios of hedge portfolios could raise doubts about the true benefits of ensuring sparsity in our framework. Therefore, we need to employ a more stringent benchmark that adheres to relevant scientific research.

One way to establish a fair benchmark is to replicate standard procedures for testing pricing anomalies while adjusting them to our framework. In that spirit, we generated out-of-sample results by designating the classic 3-factor model factors (Fama and French, 1993) as relevant - the aim is to incorporate the classic benchmark into our setup. The performance of this benchmark for all prediction windows considered is presented in table 8.

Table 8: Out-of-sample returns prediction results - benchmark

This table reports out-of-sample annualized Sharpe ratios of hedge portfolios going long-short the 10% of stocks with the highest-lowest predicted returns, setting the factors as Fama and French (1993), for different RW_{pred}.

RW_{pred}	120	180	240
3-factors Fama-French	1.57	1.64	1.49

Upon cross-examining the results presented in table 4, table 5, table 6, and table 7 alongside the new benchmark performance in table 8, we observe that surpassing the proposed benchmark is no simple task. Running the out-of-sample exercise with the 3 Fama-French factors yields gross of trading costs Sharpe ratios as high as 1.64 for $RW_{pred} = 180$. For instance, even the best results obtained

with the p-value threshold method (see table 4 and table 6), 1.49 ($RW_{shrk} = 120$, $RW_{pred} = 240$) and 1.44 ($RW_{shrk} = 240$, $RW_{pred} = 240$), failed to outperform the Fama-French based selection for all tested values of RW_{pred}.

However, several hedge portfolios managed to outperform this stricter benchmark.[16] As shown in table 9, all these portfolios were constructed under the fixed number of relevant factors criterion, and two key insights can be drawn:

- Even if the most statistically significant factor has a p-value higher than 0.10, using it provided better results than not designating any factor as relevant;
- A higher sparsity level resulted in enhanced out-of-sample predictability.

Table 9: Outperforming hedge portfolios

This table reports the portfolios presented on table 4, table 5, table 6, and table 7 that yielded better annualized Sharpe ratios than the best-performing benchmark portfolio ($RW_{pred} = 180$) - presented on table 8.

Spanning criterion	RW_{shrk}	Factor selection criterion	RW_{pred}	Mean	SD	Sharpe	Turnover
f5	180	lwst1	240	0.0135	0.0229	2.04	1.36
f5	180	lwst1	180	0.0142	0.0243	2.03	1.34
f5	120	lwst1	240	0.0131	0.0226	2.01	1.38
f5	180	lwst1	120	0.0144	0.0250	2.00	1.36
BIC	120	lwst1	240	0.0130	0.0231	1.95	1.40
f5	240	lwst3	240	0.0143	0.0259	1.91	2.18
BIC	120	lwst1	180	0.0135	0.0247	1.89	1.41
f5	120	lwst1	180	0.0131	0.0241	1.88	1.38
f5	240	lwst3	180	0.0143	0.0266	1.86	2.17
BIC	240	lwst3	180	0.0144	0.0271	1.85	2.17
BIC	240	lwst3	240	0.0140	0.0265	1.83	2.17
BIC	240	lwst3	120	0.0144	0.0277	1.80	2.27
f5	240	lwst3	120	0.0141	0.0273	1.79	2.25
BIC	180	lwst1	240	0.0121	0.0235	1.79	1.38
BIC	180	lwst1	180	0.0129	0.0252	1.78	1.36
f5	120	lwst3	240	0.0117	0.0232	1.75	2.15
BIC	180	lwst1	120	0.0130	0.0258	1.74	1.37
BIC	120	lwst1	120	0.0126	0.0254	1.72	1.39
f5	120	lwst1	120	0.0123	0.0248	1.72	1.35
f5	240	lwst1	240	0.0121	0.0246	1.71	1.40
BIC	120	lwst3	240	0.0111	0.0231	1.67	2.15
f5	180	lwst3	240	0.0120	0.0251	1.65	2.23
BIC	180	lwst3	120	0.0132	0.0279	1.64	2.12

Our findings suggest that the dimensionality issue of the zoo of factors can be satisfactorily addressed by estimating the statistical significance of factor-spanning shrinkage regressions' intercepts. Specifically, our results show that when using monthly data in the framework proposed in section 2, the most relevant factors should be selected assuming significant pre-defined sparsity, i.e., classifying fewer than 3 factors as relevant.

More interestingly, although the best out-of-sample results are generally obtained considering a combination of shorter factor selection and longer returns prediction windows, i.e., $RW_{shrk} \in [120, 180]$

[16] The appendix shows which portfolios outperformed the benchmark for all prediction rolling windows.

and $RW_{pred} \in [180, 240]$, it was possible to surpass the proposed benchmark with $RW_{shrk} = 240$ or $RW_{pred} = 120$. This indicates that our findings are likely not very susceptible to data mining, as there is no need for a specific combination of window sizes to obtain interesting out-of-sample results.

5 Conclusion

This paper introduces a novel framework designed to tackle the challenge of high-dimensionality factor-based asset pricing models. Our approach leverages shrinkage methodologies applied to regressions that span the returns of each pricing anomaly with those of all other anomalies. The versatility of our framework allows for its application across various horizons of interest, involving the segmentation of time series into distinct phases: one for controlling multicollinearity, another for factor selection, and a final phase for predicting one-period-ahead returns. Our framework is adaptable to accommodate alternative methodologies for shrinkage in the spanning regressions and returns forecasting, despite presenting results for shrinking with the LASSO and forecasting using the OLS only.

We evaluate the statistical significance of factors' relevance by examining the intercepts of spanning regressions. Lower p-values associated with these intercepts indicate that the returns of certain factors cannot be adequately explained by the returns of all other factors, suggesting that these factors carry unique information that others cannot replicate.

Furthermore, we contend that traditional factor-based asset pricing models often assume high sparsity levels arbitrarily, underscoring the potential value of ensuring a comparable scarcity of factors. This sparsity assumption can be evaluated within our framework through two key steps: first, in the spanning regressions, and second, when assessing factor relevance.

We advocate for an alternative criterion in setting the penalization parameter of shrinkage regressions, wherein it is dynamically adjusted to ensure a predefined number of jointly relevant factors are selected. This criterion proves advantageous when researchers possess prior beliefs regarding the number of factors to be considered. Furthermore, we propose enforcing a set sparsity level during the selection of relevant factors, recommending the consideration of a predetermined number of factors ranked by their intercept's p-value, regardless of any predefined statistical relevance ceiling. The outcomes presented in our study suggest that this sparsity assumption holds merit in both instances.

Moreover, we introduce a more stringent benchmark for evaluating our results. We recognize that test assets based on anomaly factors presumed to be relevant may yield a positive alpha simply as a byproduct of their construction. Thus, our proposed benchmark blends elements of our out-of-sample framework with the classic Fama-French 3 factors model, providing a fairer basis for comparison.

Applying our methodology to a comprehensive set of 80 factors drawn from the literature, along-

side the market factor, we utilized the widely-adopted LASSO technique for the spanning regressions. We fixed the rolling window for multicollinearity treatment at 240 periods and considered candidate periods of 120, 180, and 240 months for both factor selection and returns forecasting. While the results were less promising when considering only factors with associated p-values below 0.10, apart from the methodology for setting the LASSO's penalty parameter, selecting a predefined number of factors ranked according to statistical significance enabled us to surpass the stricter benchmark in various parameter combinations.

We observed that shorter windows for factor selection coupled with longer windows for return prediction generally yielded superior results. This finding is grounded in the notion that the relevance of factors should closely correspond to the present moment, while the estimation of the relationship between selected factors' returns and test assets' returns benefits from a longer horizon. However, certain combinations involving longer factor selection windows and extended time series for forecasting test assets' returns were able to outperform the proposed benchmark, indicating some robustness degree against potential data-mining concerns, and bolstering our findings' credibility.

Our framework offers researchers and practitioners a valuable tool for screening relevant pricing factors, providing an additional criterion for setting the penalization parameter and encouraging exploration into methods for ensuring sparsity within their models. Our results demonstrate that even simple regressions, in combination with properly set frameworks and criteria, can yield compelling outcomes.

As we near the conclusion, it feels as though we're stepping out of a fierce arena of financial modeling, reminiscent of the trials faced in the Hunger Games. Just as tributes compete for survival, the proposed pricing factors have engaged in intense competition within our statistical realm. Our quest has been to discern the champions – factors endowed with p-values in their favor. So, as say bid adieu to this econometric odyssey, let's celebrate the victors, for in the realm of financial modeling, the p-values were indeed in their favor.

References

H. Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

A. Belloni, V. Chernozhukov, and C. Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650, 2014.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

C. Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8:3–62, 1936.

M. J. Brennan and Y. Xia. Assessing asset pricing anomalies. *The Review of Financial Studies*, 14 (4):905–942, 2001.

M. M. Carhart. On persistence in mutual fund performance. *The Journal of finance*, 52(1):57–82, 1997.

E. Carlstein. The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *The annals of statistics*, pages 1171–1179, 1986.

J. H. Cochrane. Presidential address: Discount rates. *The Journal of finance*, 66(4):1047–1108, 2011.

E. F. Fama and K. R. French. Common risk factors in the returns on stocks and bonds. *Journal of financial economics*, 33(1):3–56, 1993.

E. F. Fama and K. R. French. Dissecting anomalies. *The Journal of Finance*, 63(4):1653–1678, 2008.

E. F. Fama and K. R. French. A five-factor asset pricing model. *Journal of financial economics*, 116 (1):1–22, 2015.

G. Feng, S. Giglio, and D. Xiu. Taming the factor zoo: A test of new factors. *The Journal of Finance*, 75(3):1327–1370, 2020.

M. Figueiredo and R. Nowak. Ordered weighted l1 regularized regression with strongly correlated covariates: Theoretical aspects. In *Artificial Intelligence and Statistics*, pages 930–938. PMLR, 2016.

A. Frazzini and L. H. Pedersen. Betting against beta. *Journal of financial economics*, 111(1):1–25, 2014.

J. Freyberger, A. Neuhierl, and M. Weber. Dissecting characteristics nonparametrically. *The Review of Financial Studies*, 33(5):2326–2377, 2020.

J. Green, J. R. Hand, and X. F. Zhang. The characteristics that provide independent information about average us monthly stock returns. *The Review of Financial Studies*, 30(12):4389–4436, 2017.

T. Gupta and B. Kelly. Factor momentum everywhere. *The Journal of Portfolio Management*, 45(3): 13–36, 2019.

P. Hall. Resampling a coverage pattern. *Stochastic processes and their applications*, 20(2):231–246, 1985.

C. R. Harvey and Y. Liu. Lucky factors. *Journal of Financial Economics*, 141(2):413–435, 2021.

S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

K. Hou, C. Xue, and L. Zhang. Digesting anomalies: An investment approach. *The Review of Financial Studies*, 28(3):650–705, 2015.

K. Hou, C. Xue, and L. Zhang. Replicating anomalies. *The Review of financial studies*, 33(5):2019–2133, 2020.

P. Houweling and J. Van Zundert. Factor investing in the corporate bond market. *Financial Analysts Journal*, 73(2):100–115, 2017.

J. Huang, J. L. Horowitz, and F. Wei. Variable selection in nonparametric additive models. *Annals of Statistics*, 38:2282–313, 2010.

T. I. Jensen, B. Kelly, and L. H. Pedersen. Is there a replication crisis in finance? *The Journal of Finance*, 78(5):2465–2518, 2023.

J. Lewellen. The cross-section of expected stock returns. critical finance review, 4 (1), 1–44, 2015.

R. K. Loh and M. Warachka. Streaks in earnings surprises and the cross-section of stock returns. *Management Science*, 58(7):1305–1321, 2012.

N. Meinshausen and P. Bühlmann. Variable selection and high-dimensional graphs with the lasso. *Annals of Statistics*, 34:1436–1462, 2006.

N. Meinshausen, L. Meier, and P. Bühlmann. P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681, 2009.

G. Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

W. F. Sharpe. The sharpe ratio. *Streetwise–the Best of the Journal of Portfolio Management*, 3: 169–185, 1998.

M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society: Series B (Methodological)*, 36(2):111–133, 1974.

C. Sun. Factor correlation and the cross section of asset returns: a correlation-robust approach. *Working paper*, 2023.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

L. Wasserman and K. Roeder. High dimensional variable selection. *Annals of statistics*, 37(5A):2178, 2009.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

6 Appendix

6.1 Statistically significant factors over time

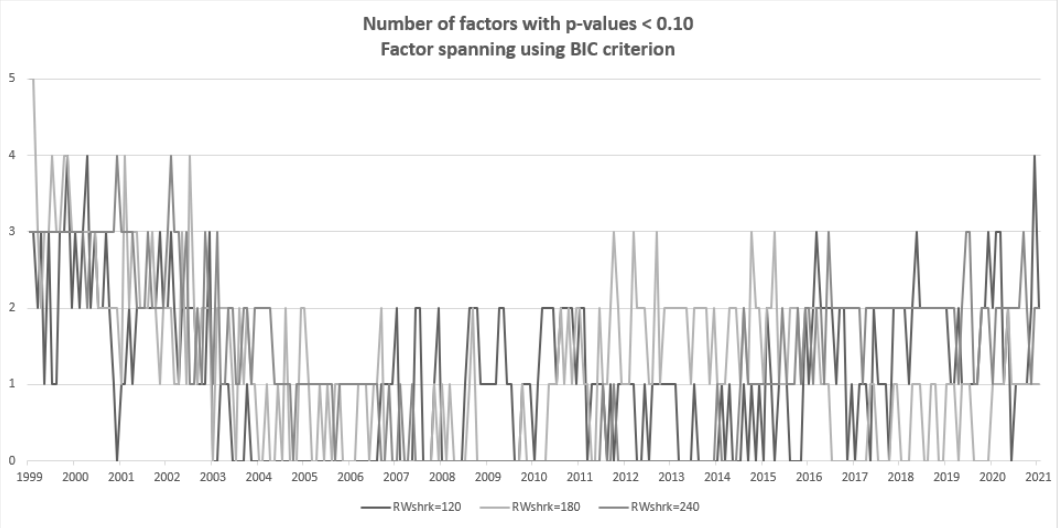

Figure 3: Pricing anomalies associated with low p-values trough time - BIC penalty

The plot reports the number of factors that presented elevated statistical significance over time, when setting shrinkage penalization parameter through BIC, and for different values for RW_{shrk}.

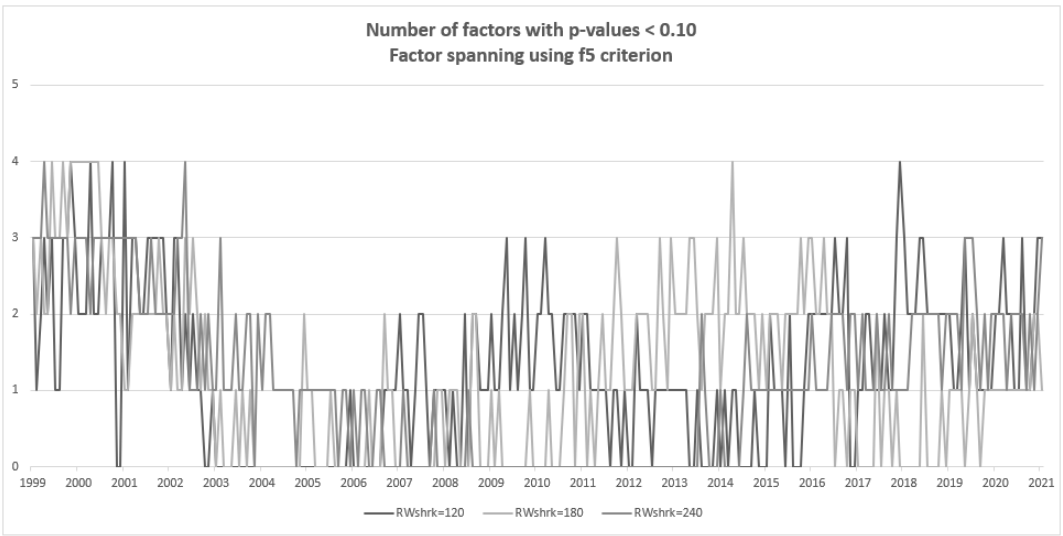

Figure 4: Pricing anomalies associated with low p-values trough time - FRC (f5) penalty

The plot reports the number of factors that presented elevated statistical significance over time, when setting shrinkage penalization parameter through (fixing 5 regressors - f5), and for different values for RW_{shrk}.

6.2 Other outperforming portfolios

Table 10: Outperforming hedge portfolios - Second-best performing RW_{pred}

This table reports the portfolios presented on table 4, table 5, table 6, and table 7 that yielded better annualized Sharpe ratios than the second-best-performing benchmark portfolio ($RW_{pred} = 240$) - presented on table 8.

Spanning criterion	RW_{shrk}	Factor selection criterion	RW_{pred}	Mean	SD	Sharpe	Turnover
all	240	lwst1	240	0.0119	0.0252	1.64	1.37
all	180	lwst3	240	0.0122	0.0261	1.63	2.10
all	180	lwst3	180	0.0129	0.0274	1.62	2.06
f5	240	lwst1	180	0.0123	0.0262	1.62	1.43
all	120	lwst3	180	0.0113	0.0245	1.61	2.20
f5	120	lwst3	180	0.0113	0.0244	1.60	2.18
f5	180	lwst3	180	0.0122	0.0265	1.60	2.20
f5	120	lwst3	120	0.0117	0.0257	1.57	2.30

Table 11: Outperforming hedge portfolios - Worst performing RW_{pred}

This table reports the portfolios presented on table 4, table 5, table 6, and table 7 that yielded better annualized Sharpe ratios than the worst-performing benchmark portfolio ($RW_{pred} = 240$) - presented on table 8.

Spanning criterion	RW_{shrk}	Factor selection criterion	RW_{pred}	Mean	SD	Sharpe	Turnover
all	240	lwst1	180	0.0121	0.0268	1.56	1.41
f5	180	lwst3	120	0.0125	0.0277	1.56	2.28
all	120	lwst3	120	0.0116	0.0260	1.54	2.30
f5	240	lwst1	120	0.0120	0.0270	1.54	1.50
all	240	lwst1	120	0.0120	0.0276	1.51	1.48