MASTER THESIS

Option Factor Timing

Author

OSKAR VON REICHENBACH*
Master of Science in Finance (MFin), Stockholm School of Economics

Supervisors

PROF. TOBIAS SICHERT, PH.D.
Swedish House of Finance, Stockholm School of Economics

PROF. DR. ROLAND FÜSS
Swiss Institute of Banking and Finance, University of St. Gallen

May 3, 2026

*42751@student.hhs.se

Abstract

I demonstrate that factor return patterns are predictable in the time series of option factors using a combination of PCA and linear and nonlinear machine learning models. The principal components that capture volatility-, and stock-quality-related patterns are predictable and achieve out-of-sample R^2 of up to 33%. This predictability translates into economically meaningful gains; factor timing portfolios using ensemble methods yield annualized returns of 17–20%, compared to 4% for an equal-weighted benchmark. However, factor timing results in lower risk-adjusted returns, primarily due to difficulties in estimating the covariance matrix. When accounting for transaction costs, factor timing results in negative returns. The results are robust across various specifications, including alternative estimators, portfolio construction, and separate analyses of puts and calls.

JEL classification: G10, G11, G12, G13

Keywords: option factors, factor timing, return forecasts, principal component analysis, empirical asset pricing, machine learning

Contents

1 Introduction

Option returns exhibit predictable cross-sectional patterns. Recent research has identified numerous option factors that explain these patterns, creating an option "factor zoo" similar to equity markets (Zhan, Han, Cao, & Tong, 2022; Cochrane, 2011). The question is whether these fluctuations can be anticipated. This essentially boils down to the issue of factor timing. Factor timing strategies have proven successful for equity factors, motivating their application to option markets (Haddad, Kozak, & Santosh, 2020; Kagkadis, Nolte, Nolte, & Vasilas, 2024).

I examine whether option factor returns are predictable through several Machine Learning (ML) methods and predicting them translates into economically significant gains. Using principle component analysis (PCA) to reduce the dimensionality of 28 option factors, I find that six principal components (PCs) capture 69% of the linear variance in option factor returns. Linear and nonlinear ML models achieve statistically significant forecasting performance for the first and second PC, with ensemble methods producing out-of-sample R^2 values of up to 33% for the first PC, capturing mainly volatility- and stock-quality-related patterns, and 15% for the second PC, representing equity-side frictions and limits-to-arbitrage. From individual models Ridge, Lasso, Elastic Net, XGBoost, and Dart perform best. Linear and nonlinear models yield similar results. Some linear models outperform certain nonlinear models, and vice versa. Hence, employing nonlinear models does not provide significant improvements in predicting PCs.

The forecasting results translate into substantial economic gains. Factor timing portfolios constructed from ensembles of ML predictions generate annualized returns between 17% and 20%, compared to 4% for an equally weighted benchmark representing a static factor investment strategy. Yet, after adjusting for risk, the equally weighted factor portfolio achieves the highest Sharpe Ratio, about 2.7, which significantly outperforms most factor timing strategies at a 95% confidence level. These results imply that, although ML models can predict option factor returns, difficulties in estimating high-dimensional covariance matrices limit the effectiveness of Sharpe Ratio optimization for factor timing strategies.

My findings are generally robust across multiple specifications. These include direct and indirect factor weights estimation, exclusion of the Great Financial Crisis (GFC), separate analysis of calls and puts, factor selection approaches, and various methods of incorporating option characteristics as explanatory variables. Never-

theless, the results deteriorate under expanding window estimation and when accounting for transaction costs. When accounting for transaction costs, the optimal factor return portfolio achieves negative returns, although the losses are smaller than those of the equal-weighted portfolio. Excluding forecasted negative returns does not solve this problem. I discuss several strategies for mitigating transaction costs that could push returns into positive territory.

These results contribute to the growing body of literature on option factors and factor timing. My analysis reveals that predictable components exist, particularly volatility-related ones. I further show that when using PCA for dimensionality reduction, the benefit of applying nonlinear ML models for option factor prediction is limited. Finally, I illustrate that successful implementation requires careful attention to portfolio construction and transaction costs. Challenges remain in translating statistically significant forecasting performance into risk-adjusted economic profits.

After I discuss the related literature in Section 2, I present the data sources and their properties in Section 3. The discussion of key methods used for dimensionality reduction, forecasting models, and forecasting assessment is presented in Section 4. Finally, Section 5 discusses the factor dimensions, forecasting performance and its economic impact, as well as several robustness checks.

2 Related literature

As the notion of both factor timing as well as option factors is a more recent topic in asset pricing, the research has mainly focussed on both subjects separately. My thesis expands current research in both areas. In the following, I discuss related literature and show how my research is related.

Option factors. Recent research has uncovered a wide array of equity-option risk factors. Under the classical option theory of Black and Scholes (1973) options are redundant. This means that creating a delta-neutral option position should generate the risk-free rate. A delta-hedge combines a stock option with a short position of the hedge, to neutralize the portfolio to local changes of the stock price (Coval & Shumway, 2001).

This notion is challenged by empirical findings of Coval and Shumway (2001). The authors find negative expected returns for delta-neutral at-the-money straddle po-

sitions. They suggest that option-specific risk factors could drive this observation. Bakshi and Kapadia (2003) find direct evidence of the volatility risk premium embedded in options. The negative return of delta-hedged S&P 500 index calls implies a negative market volatility risk premium: option sellers earn excess returns as compensation for bearing volatility fluctuations.

Since then, many studies have identified specific option factors. Using these factors, one can construct models that predict the expected return of delta-hedged options.[1] In recent years, the number of delta-hedged option factors has been steadily growing. Although it is not yet reached the size of the stock factor "zoo", the number of factors has grown to a considerable amount (Feng, Giglio, & Xiu, 2020). I provide an overview and cluster the option factors in Appendix A.1. My thesis examines 28 well-established option return factors, summarized in Table B1.

Factor timing. My thesis explores factor timing in U.S. stock option markets, an area that has received little attention. To my knowledge, no study has systematically tested the timing of option factors. I also compare the use of linear and nonlinear ML models in this setting. While these methods have been applied to equity factor timing, they have not been used for option factors. However, I build heavily on recent advancements in factor timing for stock factors. Although Asness, Chandra, Ilmanen, and Israel (2017) highlight significant practical challenges of factor timing, such as the weak predictive power of value spreads and the dilution of timing benefits due to diversification, recent studies have made notable progress.

First and foremost, my thesis relates to Kagkadis et al. (2024), who show that factor timing is possible when using combinations of dimensionality reduction techniques. The study employs PCA and Lasso, among others, to predict factor returns. The key innovation lies in making use of the factor-specific characteristics. In general, dynamic factor timing improves risk-adjusted performance compared to static allocations, such as static multi-factor portfolios. Furthermore, they discovered significant heterogeneity across economic states and stock factors. Their results are generally robust across model specifications and markets. I base my assessment of U.S. stock option factors on their methodology, but expand it to include other linear and nonlinear models.

[1] Christoffersen, Fournier, and Jacobs (2018) also explore the factor structure of options. Instead of predicting delta-hedged option returns, they examine the relationship between the underlying asset and the equity option. Using a factor model, they find that firms with higher market betas have higher implied volatilities, steeper moneyness slopes, and a term structure that covaries more with the market.

Haddad et al. (2020) also reduce dimensionality through PCA to confirm that stock factors are predictable, with some PCs achieving up to four times higher out-of-sample R^2 than aggregate market returns. This leads to a higher risk-adjusted return than market timing and static factor investing. The authors also theoretically demonstrate that the factor timing portfolio is equivalent to the stochastic discount factor (SDF). They discover that an SDF incorporating factor timing increases in volatility, implying that it captures more time-varying risk.[2] Haddad et al. (2020) also lay the theoretical foundation of factor timing. The authors outline two key assumptions that I mention in Appendix A.2.

In a recent working paper, Neuhierl, Randl, Reschenhofer, and Zechner (2023) test the factor timing of 300 U.S. stock factors. They find that past factor returns and volatility are the most reliable predictors of future factor returns. Using partial least squares (PLS) for dimensionality reduction, they observe on average an annual increase of returns of 2% when timing factors. In a related matter, Ma, Liao, and Jiang (2023) use a deep learning approach to successfully time factors in the Chinese stock market. Lehnherr, Mehta, and Nagel (2025) emphasize the usage of specific shrinkage methodologies that are required when using many predictors for factor timing. Similarly, I discuss the necessity of using shrinkage and dimensionality reduction when working with high-dimensional data, particularly given that my data is only available monthly.

Recent work has also been put into the momentum of factor returns: Gupta and Kelly (2019) find that factors can be timed through their momentum. In essence, asset pricing factors exhibit positive autocorrelation. In addition, Ehsani and Linnainmaa (2022) find that when applying a momentum strategy, momentum does not represent a distinct risk factor; rather, it dynamically times other risk factors. Käfer, Mörke, and Wiest (2025) test their hypothesis and confirm it for options. They also demonstrate that factor momentum is reflected in market expectations using option implied returns. I deploy several momentum predictors and characteristics and find that they have a high level of importance, especially for linear models such as Lasso or PLS.

My work extends factor timing research beyond equity markets into options. It also clarifies the practical challenges of applying these techniques, particularly the

[2]Additionally, Haddad et al. (2020) find that loadings on factors such as value and size are procyclical, while momentum is countercyclical. This lends more credence to the notion of Moreira and Muir (2017) that taking less risk exposure when past volatility was high involves dynamically adjusting factor exposures.

role of covariance matrix estimation noise in affecting Sharpe Ratio performance. My research identifies both the potential and the methodological requirements for successful factor timing in option markets. The key change is acknowledging that while predictability research exists in options (like An, Ang, Bali, and Cakici (2014)), my work is the first to systematically apply factor timing methodology specifically to option factor returns.

Methods. My thesis assesses and applies several ML and econometric methods. On the one hand, these methods are necessary for forecasting the target variable, such as factor or equity option returns. For example, predicting returns in the simplest form requires a regression or prediction algorithm. On the other hand, these methods are necessary for making inferences about the drivers and the nature of the relationships. For example, determining whether variables are eliminated in a Lasso regression can provide insight into whether the variable is necessary for the forecast.

Recent advancements in the empirical asset pricing literature have been made, especially in the domain of predicting asset returns using ML methods. With the large amount of factors on one side, and the large amount of predictors (Goyal, Welch, & Zafirov, 2024) on the other side, conventional econometric tools reach a limit. B. Kelly and Xiu (2023) point out that ML methods can be used in this case as they enable the usage of many predictors without high risk of overfitting, like in ordinary least squares (OLS).

In terms of methodology, my thesis is closely related to Bali, Beckmeyer, Mörke, and Weigert (2023). The authors compare linear and nonlinear ML methods to predict the cross section of delta-hedged option returns. They find that nonlinear models, such as gradient boosters or neural networks, achieve higher forecasting performance and risk-adjusted returns than linear models, such as Lasso or PLS. Similarly, Goyenko and Zhang (2020) uses similarly linear and nonlinear methods to predict both the cross-section of stocks and options. They show that ML and large set of predictors results in superior performance. They further find that option-based predictors have highest predictive power for both stock and option returns.[3] Similar to Bali et al. (2023), I compare linear and nonlinear forecasting

[3]Several studies confirm that ML methods outperform regular regressions in predicting asset returns. Gu, Kelly, and Xiu (2020) confirm this by finding that ML methods in general increase the forecasting performance of stock returns. Nonlinear methods are the best performing, especially neural networks and tree-based algorithms. Also, they provide tools to assess the quality of the

models. Yet, my findings differ from theirs. I show that nonlinear methods do not deliver measurable improvements over linear ones. This holds when combined with linear PCA for predicting monthly factor returns.

ML methods make it possible to use large sets of return predictors. This requires a different way of thinking about risk factors. The classical approach of Fama and French (1993) emphasizes using only a few predictors. This is at odds with recent findings. Similar to other studies, Didisheim, Ke, Kelly, and Malamud (2024) empirically find that the out-of-sample performance of factor pricing models increases with the number of factors. The authors then theoretically demonstrate that, if the underlying factor structure is not too concentrated (i.e., if there is no concentrated eigenvalue distribution of factors), adding factors to the model is beneficial. Martin and Nagel (2022) theoretically analyze the relationship between market efficiency and the predictability of asset returns using big data. They show that in-sample return predictability can emerge due to high-dimensional learning constraints while market efficiency still holds. Therefore, out-of-sample tests have more economic significance and are better suited to detecting true return predictability. In my thesis, I only use out-of-sample measures of forecasting performance. In contrast, Murray, Xia, and Xiao (2024) model neural networks that use historical stock performance to predict stock returns and find significant prediction performance, even out-of-sample. This finding calls into question the efficient market hypothesis which states that past information cannot be used to predict future returns because prices already incorporate it.

I use PCA to reduce the dimensions of the dependent variable. This enables me to perform an OLS regression to predict returns. Following Kagkadis et al. (2024) and Haddad et al. (2020), I also apply PCA to reduce the dimensions of my predictors. Several recent papers reduce the dimensions of their estimation problems. This is especially important when using many predictors. B. T. Kelly et al. (2019) design an instrumented PCA approach.[4] Alternatively, some papers reduce dimensional-

predictions, such as out-of-sample R^2. Based on work from B. T. Kelly, Pruitt, and Su (2019), Gu, Kelly, and Xiu (2021) model factor exposures through autoencoder neural networks. This allows to incorporate both factors (as nonlinear functions of covariates) and returns. Grammig, Hanenberg, Schlag, and Sönksen (2025) find that ML perform better over one-year horizon forecasts. Selecting test assets through the no-arbirage condition, L. Chen, Pelger, and Zhu (2024) use a non-parametric deep learning model to estimate the SDF of assets. The paper includes states of the economy in the prediction.

[4]Büchner and Kelly (2022) find three latent factors that explain 85% of the variation in S&P 500 option returns. Based on the same framework, Goyal and Saretto (2025) construct a factor model of equity option returns. The authors achieve an alpha of close to zero and test different trading

ity by constructing sparse models through Lasso (Freyberger, Neuhierl, & Weber, 2020; Messmer & Audrino, 2022; Shafaati, Chance, & Brooks, 2023).

Dimensionality reduction is not always optimal. This is especially true when the underlying model does not rely on sparsity. Käfer, Moerke, Weigert, and Wiest (2025) apply the Bayesian SDF of Bryzgalova, Huang, and Julliard (2023) to price options. This framework was originally developed for stock returns. It offers a simple and scalable solution for estimating linear asset pricing models in high-dimensional settings. The framework remains robust even under weak identification. Their evidence points to a dense true SDF. This suggests that many factors may be relevant simultaneously. In contrast to approaches that aggressively shrink the factor set, my study takes a different approach. I retain a broad cross-section of option factors. I then apply ML methods to capture time-series predictability in their returns. I find that feature importance varies significantly across different models. No single predictor consistently dominates factor timing predictions. This variability supports keeping a large predictor set, consistent with findings from Käfer, Moerke, et al. (2025). Although I use many factors, I do not directly address the dense SDF found by Käfer, Moerke, et al. (2025). I employ dimensionality reduction and regularization to control complexity. Future research could directly address dense SDFs by integrating Bayesian SDF frameworks into factor timing models.

3 Data

Since I use ML methods such as cross-validation, model-internal regularization, and dimensionality reduction techniques to control overfitting, I incorporate all possibly beneficial predictors in my estimation. B. Kelly and Xiu (2023) highlight the benefits of this approach. In this section, I present both the predicted and predictor variables used in the analysis.

3.1 Option returns and factors

My main data set includes monthly delta-hedged option returns and factor values. I use this data for both the dependent variable (factor return series) as well as for the predictor set (factor characteristics). The data source is proprietary data from

strategies with which they have mothly realized returns of 80 basis points. The working paper of Horenstein, Vasquez, and Xiao (2023) constructs a comparable latent factor model and achieves similar results.

OptionMetrics IvyDB, which includes historical prices for all exchange-traded U.S. single equity options. The construction of delta-hedged returns follows Bakshi and Kapadia (2003) and spans data from 1996 to 2022.[5] Table 1 includes summary statistics on the delta-hedged returns.

Variable	Mean	SD	10^{th} pct.	Median	90^{th} pct.
Panel A: Calls					
Option return (daily delta-hedged)	-0.001	0.06	-0.05	-0.01	0.05
Dollar open interest	1824.863	9234.91	8.75	154.88	3342.34
Delta	0.537	0.11	0.39	0.54	0.68
Moneyness (K/S)	1.002	0.05	0.94	1.00	1.06
Time to maturity (in days)	49.652	2.08	46.00	50.00	52.00
Market capitalization	9949.075	42512.78	338.50	1891.97	18463.03
Panel B: Puts					
Option return (daily delta-hedged)	-0.002	0.05	-0.04	-0.01	0.04
Dollar open interest	1327.671	6899.64	7.31	107.21	2336.25
Delta	-0.457	0.12	-0.61	-0.45	-0.31
Moneyness (K/S)	0.998	0.05	0.94	1.00	1.06
Time to maturity (in days)	49.637	2.08	46.00	50.00	52.00
Market capitalization	11004.166	45040.10	379.62	2145.63	20994.87

Table 1: Summary stats on delta-hedged returns. Call and put options exhibit similar patterns, though there are some key differences. Both have slightly negative average delta-hedged returns: calls are at -0.001, while puts are lower at -0.002. Calls exhibit greater variation in return s than pu ts. Call deltas average 0.54, whereas put deltas are negative and farther from zero. Both option types cluster around at-the-money strikes with similar 50-day maturities. On average, calls attract higher dollar open interest. Market cap distributions are highly skewed for both types. In general, calls have higher return volatility and greater trading interest.

The factor values include both traded and non-traded factors. Tradable factors are used to construct factor returns (shown in Table B1). Non-tradable factors that I use for prediction are listed in Appendix B.3.

Factor returns. To construct factor portfolio returns, I form long-short portfolios based on the first and tenth deciles of each factor's cross-sectional distribution. The direction of the position (long decile 10 and short decile 1, or vice versa) is determined by the cumulative performance of delta-hedged returns since January 1996. Finally, I scale the factor returns.

Figure 1 shows the monthly factor returns from 1996 to 2022. Overall, the returns appear to be positive on average, though there is noticeable variation in volatility

[5] Although the factor dataset spans data from 1996 to 2022, I only use data up to the end of 2019 for the statistical and economic analysis of factor timing, as not all predictor variables are available for 2020 onwards.

over time. Periods such as the early 2000s and the 2008 financial crisis exhibit heightened volatility and wider dispersion in factor portfolio performance. These fluctuations likely reflect the turbulence experienced in the broader market during the dot-com bubble, the global financial crisis and the pandemic. In contrast, the periods between these events show more stable and less volatile return patterns.

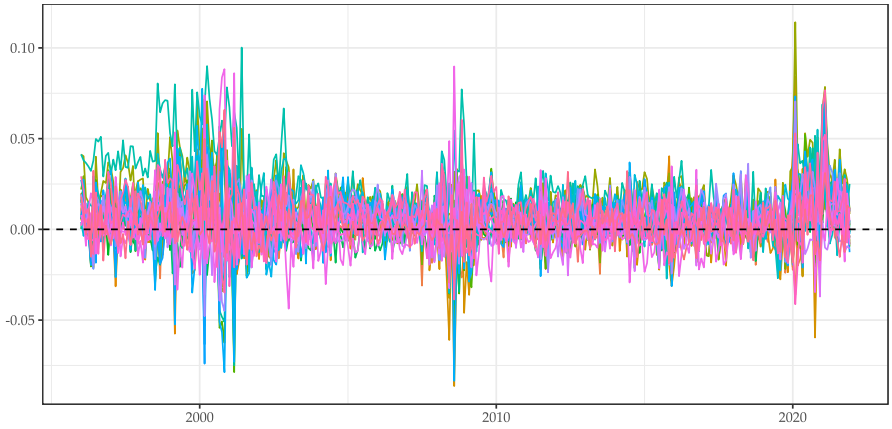


Figure 1: Monthly returns of factor portfolios; each color represents one factor portfolio. Although factor returns are generally positive, there are differences over time as well as between factors. For example, more extreme values occur more frequently during times of financial distress, such as the GFC. Additionally, some factors tend to perform better than others during times of crisis.

This is observation confirmed by summary statistics shown in Table *B*2. The monthly mean return is positive and significantly different from zero for almost all factor portfolios. Furthermore, I find that the return series of each factor portfolio does not follow a normal distribution, but is leptokurtotic. Depending on the factor portfolio, I find both positive and negative skewness. Thus, factor returns have asymmetries and tail risks.

Characteristic weightings. I construct the set of characteristics in two stages. First, I use the factor values directly as characteristics. Specifically, I compute the average characteristic value difference between the first and tenth deciles. Second, I derive additional characteristics based on these mean factor portfolio values, capturing both momentum and cross-sectional spread effects. Table 2 details the construction of these derived characteristics.

Using the raw characteristic values - including both the original factor values per portfolio and the derived momentum and spread characteristics - I construct factor weightings for each factor portfolio. I do this by scaling raw values to range

Characteristic	Source	Implementation
mom1	Käfer, Mörke, and Wiest (2025)	Factor returns of $t-1$
mom2	Käfer, Mörke, and Wiest (2025)	Factor returns of $t-6$
mom3	Käfer, Mörke, and Wiest (2025)	Mean factor returns from $t-2$ to $t-12$
mom4	Käfer, Mörke, and Wiest (2025)	Mean factor returns from $t-13$ to $t-60$
spread1	Neuhierl et al. (2023)	Raw decile-spread $S_t = \bar{d}_{10,t} - \bar{d}_{1,t}$, then $spread1 = \frac{S_t - \overline{S}_{1:t}}{\sigma_{S_{1:t}}}$.
spread2	Neuhierl et al. (2023)	Raw decile-spread S_t, then $spread2 = \frac{S_t - \overline{S}_{(t-5):t}}{\sigma_{S_{(t-5):t}}}$.

Table 2: Additional characteristics added to the C matrix. The spread-based signal from Neuhierl et al. (2023) originates from applications to individual stock characteristics. Käfer, Mörke, and Wiest (2025) find various forms of persistence in factor returns across different horizons. The spread characteristics that detect persistence in factor returns across different horizons have been proven to help for stocks, motivating their application to option factors. (Neuhierl et al., 2023)

between -1 and 1 in each month. By design, a factor's weight on itself is either -1 or 1, depending on whether the strategy takes a short position in the first decile and a long position in the tenth, or the reverse.

3.2 Other predictors

Next to characteristic weights, I use other variables for the prediction of factor returns: Non-tradable option factors, other economic predictors, and momentum variables of PC returns.

Non-traded factors. 15 non-traded factors listed in Appendix B.3 are used as predictors. Non-traded factors span intermediary capital risk to macroeconomic and sentiment indicators, covering mainly uncertainty measures, volatility and tail-risk metrics, liquidity conditions, credit and term structure spreads, and overall financial stress in the economy.

Further economic indices. I use additional economic indicators for the prediction.

- *Borrowing costs*: The risk-free rate, term premium, and credit risk directly affect option prices, and their SDFs. I include U.S. treasury yields for three months and 30 years. The data are downloaded from FRED website. These

are explicit yields, while the factor form is contained in the non-traded factors. Additionally, I use corporate bonds yields for AAA and BAA rated bonds. These data are taken from FRED. I use the actual bond yields instead of the factor innovations in the DEF variable that are already contained in the non-traded factors.

- *Macroeconomics*: I capture macroeconomic uncertainty using the inflation rate, measured by the consumer price index (CPI), downloaded from the website of the U.S. Bureau of Labor Statistics. To account for consumption risk, I use the consumption–wealth ratio (Cay) from Lettau and Ludvigson (2001), downloaded from his website.

- *Investor sentiment*: I add the non-orthogonalized sentiment of Baker and Wurgler (2006), which reflects irrational exuberance or pessimism (the orthogonalized is already contained in non-traded factors). The data are downloaded from Jeffrey Wurgler's website. To capture both bearish sentiment and short-sale constraints, I also use the short interest index of Rapach, Ringgenberg, and Zhou (2016), downloaded from Matthew Ringgenberg's website. In order to prevent look-ahead bias, I employ out-of-sample weights. Goyal et al. (2024) find that it maintains a good out-of-sample performance to predict the equity premium.[6]

- *Equity risk factors*: Although the Fama and French (2015) five factor models are explanatory and generally not used for prediction, I include them because option returns could load on stock factors. The data are downloaded from Kennth R. French's website. I separately include the cyclically adjusted price earnings ratio (CAPE), downloaded from Robert Shiller's data webpage.

Notice that Goyal et al. (2024) show that most non-traded factors and economic indices presented above (e.g., volatility- or sentiment-based predictors) do not exhibit a positive predictive performance in forecasting the equity risk premium. I use them nevertheless because my focus is on options, where they could still prove effective.

[6]Goyal et al. (2024) show that other predictors, such as a combination of technical indicators presented by Neely, Rapach, Tu, and Zhou (2014), continue to achieve good out-of-sample predictive performance of the equity premium. However, due to data availability, I only include the short interest index.

PC momentum. Käfer, Mörke, and Wiest (2025) find significant positive factor return autocorrelation. Hence, I construct three mutually exclusive momentum variables for the PC prediction. They are distinct from *mom*1 to *mom*4 in that they are direct momentum variables of the principle components. For each PC, they are constructed as follows:

- *momentum1* captures the return of each PC in the previous month. For each time t, the momentum is simply the return at $t-1$. This variable captures monthly factor momentum, where Käfer, Mörke, and Wiest (2025) find that positive autocorrelation of factor returns is the main source for short-term momentum.
- *momentum212* represents the average momentum over the last year. For each time t, it takes the cumulative returns from the 11 months preceding the last month (i.e., from $t-12$ to $t-2$) for each PC. This avoids the most recent return $(t-1)$ to reduce noise and captures more persistent trends that exclude short-term momentum effects. Käfer, Mörke, and Wiest (2025) find that both persistent mean returns and autocorrelation drive annual momentum.
- *momentum1360* captures long-term momentum by taking the cumulative return from months $t-60$ to $t-13$. It thus uses data from more than one year ago up to about five years in the past, while deliberately excluding the most recent year (months $t-12$ to $t-1$). This isolates longer-term patterns and helps detect persistent structural return behavior. Käfer, Mörke, and Wiest (2025) identify mean return persistence as key factor.

Find a simplified process chart displaying the data pipeline in Figure B1.

4 Methodology

This section outlines the methodology for forecasting factor returns and evaluating their predictive performance.

4.1 Dimensionality reduction

My dataset is high-dimensional. The dependent variables, factor returns, include 28 series. The dimensionality of the independent variables is also high. The resulting ratio of variables to observations is high, particularly since I only have monthly data with a small number of observations. Therefore, dimensionality reduction is

essential on both sides. Following Haddad et al. (2020) and Kagkadis et al. (2024), I apply various PCA approaches to address this problem.

At the highest level, PCA transforms variables into uncorrelated variables through linear combinations such that most of the variance is then contained within the first PCs (Jolliffe, 2002). See Appendix C.1 for more mathematical background.

PCA in the dependent variable My analysis is based on 28 individual option factor returns. To reduce the complexity in the dependent variable, I employ dimensionality reduction through PCA.

The selection of PCs for analysis is crucial. Instead of relying solely on explained variance, I use the Parallel Analysis approach from Horn (1965) to identify significant PCs.[7] Using 1000 bootstrapping iterations, I find that the first six PCs are significant. Figure 2 compares the explained variance of the simulated data against the true data. I use the entire dataset to compute the covariance matrix (from 1996 to the end of 2019). This result conforms with the analysis of Haddad et al. (2020) and Kagkadis et al. (2024), who similarly use five PCs for factor timing with equities.

PCA in the independent variables. I reduce the dimensionality of the factor loadings of each PC by applying the method of Kagkadis et al. (2024). I construct the characteristics matrix by following the steps below:

1. Build a factor portfolio for each factor. I go long the first (tenth) decile and go short the tenth (first) decile, depending on cumulative delta-hedged option return of the deciles.
2. For each portfolio, I then create the average factor value for every factor (e.g., for the factor portfolio, I take for each factor its mean of decile 10 minus the mean value of decile 1).
3. Finally, I scale the values for each factor.

[7]Parallel Analysis compares the explained variance of each PC to the average explained variance of a randomly generated dataset, capturing noise. A PC is considered significant if its explained variance exceeds that of the random data. Unlike the method of Horn (1965), which uses purely random data points assuming a distribution, I adopt the nonparametric permutation-based approach from Buja and Eyuboglu (1992). I shuffle the existing factor return series independently, preserving the noise while removing time-dependent relationships. This avoids assumptions about the data distribution.

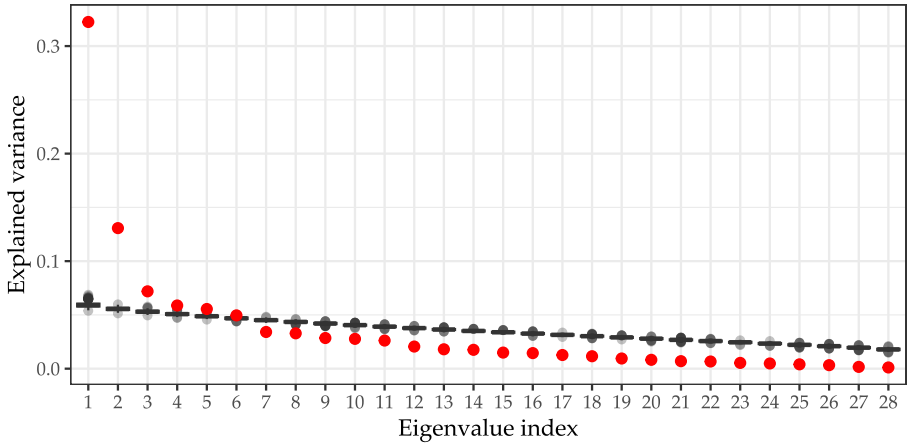


Figure 2: The explained variance of the true PCs (red) and Parallel Analysis PCs (black). The boxplots indicate that the sixth true PC exceeds the 75th percentile of the variance explained by the artificial PCs. Hence, the first six PCs are significant.

This procedure is then repeated for each time step $t = (1, \ldots, T)$. This gives us the weight caracteristics matrix $C \in \mathbb{R}^{T \times M \times N}$ where M is the number of characteristics and N is the number of factor portfolios. To retrieve the factor loadings of each PCA, I multiply the weigths matrix with the characteristics matrix:

$$H_{i,t} = w_{i,t}^{\top} C_t, \tag{1}$$

where $w_{i,t} \in \mathbb{R}^N$ is a vector of factor weights per eigenvector and $C_t \in \mathbb{R}^{N \times M}$ is the characteristics matrix. Repeating this for each $t = (1, \ldots, T)$ gives $H_i \in \mathbb{R}^{T \times M}$, a matrix containing the factor loadings for PC i.

Kagkadis et al. (2024) propose to further reduce the dimensionality of the factor loadings to reduce multicollinearity and improve predictability. I use PCA to reduce the dimensionality:

$$X_i = H_i Q_{t,i}, \tag{2}$$

where $Q_{t,i} \in \mathbb{R}^{M \times M}$ is a matrix of eigenvectors estimated at time t and sorted by corresponding eigenvalues. I estimate $Q_{t,i}$ by the eigendecomposition of $Var(H_i)$. The resulting $X_i \in \mathbb{R}^{T \times M}$ is a matrix of linear combination of the underlying characteristics. The practical advantage of doing this final transformation is that this eliminates the correlation between the characteristics, potentially enhancing the

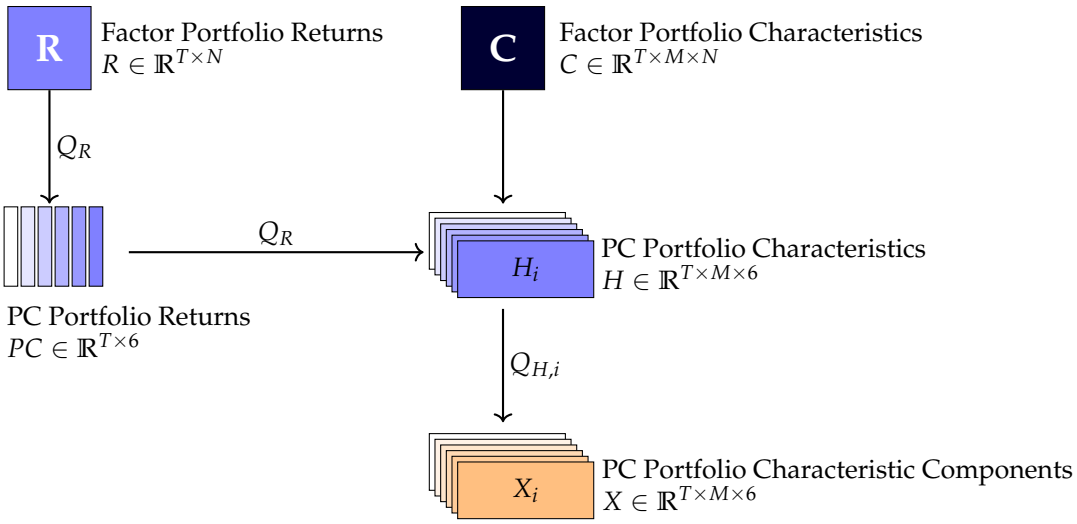


Figure 3: Overview of the transformation from factor portfolio returns and characteristics to PC-based components using weights and dimensionality reduction; based on Kagkadis et al. (2024).

overall forecasting performance (Kagkadis et al., 2024).

Finally, I combine the lagged characteristic scores in X_i with the scaled predictive variables introduced in Section 3.2. The resulting data forms the set of features used in the forecasting models.

4.2 Models

I use different models to forecast the PCA returns. I use both linear and nonlinear ML models for forecasting to evaluate whether priority should be given to linear or nonlinear approaches.

Following Bali et al. (2023), I use Lasso, Ridge, Elastic Net (EN), and PLS as linear models. To test whether simple autocorrelation can better explain movements in the PCs compared to factor weights or other predictors, I additionally use an ARMA model. The lag orders are selected using the corrected Akaike Information Criterion (Hyndman & Khandakar, 2008).

As proposed by Bates and Granger (1969), I combine all linear forecasts into one equally-weighted average forecast ensemble:

$$\hat{y}_{t+1}^{\text{L-ens}} = \frac{1}{4}(\hat{y}_{t+1}^{\text{Lasso}} + \hat{y}_{t+1}^{\text{Ridge}} + \hat{y}_{t+1}^{\text{EN}} + \hat{y}_{t+1}^{\text{PLS}}), \tag{3}$$

where the different $\hat{y}_{t+1}$ represent the respective one-month ahead forecast. Timmermann (2018) discusses several benefits of using forecast combinations. It is often not clear which model will in future perform best, especially in instable forecasting environments and when there is no clear winner among the models. Also, diversification of model uncertainty helps improving the overall forecasting performance.

Generally, linear models are able to achieve good variable reduction and dimensionality reduction. This allows them to be good in using limited data on many different explanatory variables. Also, they can deal well in maintaining interpretability. When using a linear models, for example Lasso or Ridge, coefficient values can explain how much one variable contributes to the forecasting of the dependent variable. However, linear models lack in the ability to account for non-linear patterns in the data. For example, when taking a classical PLS regression, the regression does only focus on the linear weightes of linear latent factors. As they are simple eigenvectors, they do not capture nonlinear comovements of variables, but just correlation.[8]

I find that the PC returns contain nonlinearities as shown by the Ramsey (1969) Reset test. Find results in Table 3. Additionally, B. Kelly and Xiu (2023) point out, that nonlinear ML tools can not only help to capture non-linear patterns, but also make it possible to combine multiple data sources and types.

To account for nonlinearities and interactions among the predictors, I supplement the linear models with three nonlinear ML models: Random Forests, XGBoost and Dart. These models are well-suited to high-dimensional problems and complex functional forms, as they do not make any linearity assumptions. Through their tree-based structure, they can also capture interactions and intricate relationships between predictive variables (Bali et al., 2023; Zhan et al., 2022).

As with the linear ensemble, I create an equally weighted nonlinear ensemble forecast.

$$\hat{y}_{t+1}^{\text{NL-ens}} = \frac{1}{3}(\hat{y}_{t+1}^{\text{XGB}} + \hat{y}_{t+1}^{\text{DART}} + \hat{y}_{t+1}^{\text{RF}}), \tag{4}$$

where each component represents a one-step-ahead forecast generated by the respective nonlinear model.

[8]There are efforts to incorporate nonlinear patterns within PLS, including kernel-based PLS, spline-based PLS, and neural network augmented PLS models. These are not be part of this thesis (Rosipal & Krämer, 2006).

PC	Ramsey (1969) Reset	Harvey and Collier (1977)	Utts (1982) Rainbow
1	0.00006	0.93987	0.00035
2	0.00003	0.43436	0.00000
3	0.08087	0.01647	0.00372
4	0.00504	0.36060	0.00118
5	0.32325	0.36764	0.05167
6	0.00743	0.73174	0.00147

Table 3: p-values from linear specification tests for the first six principal components (PCs) across the entire sample. The Ramsey (1969) Reset test is used to test for general functional misspecification for second- and third-level polynomials. The small p-values (mostly below 0.05) show that nonlinear patterns (in the form of second- and third-order polynomials) are highly likely. The Harvey and Collier (1977) test checks if the mean of the residuals is significantly different from zero (which would mean that there is a convex or concave structure in the residuals). We can see that we cannot reject the null hypothesis of correct model specification, except for PC3. Hence, this contradicts the results from the Ramsey (1969) test. The Utts (1982) Rainbow test verifies if there are discrepancies between the central subsample near the median and the entire sample. We reject the null hypothesis for almost all PCs; therefore, there appear to be structural breaks and linear models seem to be misspecified. Overall, these test results suggest that nonlinear patterns are likely present and that nonlinear models could be beneficial.

To account for model drift and improve robustness, all non-linear models are retuned on a rolling basis. Specifically, the hyperparameters of each model are reoptimised every twelve months using cross-validation. The hyperparameter spaces explored for tuning are listed in Appendix C.4. This approach ensures that the models adapt to changes in the data-generating process, while avoiding overfitting to short-term noise. I choose to tune only every twelve months due to performance considerations.

Nonlinear models offer considerable flexibility and often yield higher predictive accuracy in complex settings. Yet, this comes at the cost of interpretability and transparency, as their functional form is data-driven and not easily expressed in closed form. Nevertheless, Bali et al. (2023) find that nonlinear have superior performance than nonlinear models in predicting delta-hedged option returns. This is why I test the performance of nonlinear models in predicting option factor patterns. Find more in-depth information about the used linear and nonlinear models in Appendix C.2.

4.3 Forecasting

Training window. I estimate each model on a rolling window of 60 months. A rolling window lets the estimation sample "refresh" every month, so the model

can adapt to structural breaks and de-emphasise stale information. The trade-off is that, with only five years of data in each window, I work with relatively few observations compared with the large set of predictors, risking overfitting. To control this risk I combine (i) cross-validation, (ii) dimensionality-reduction (PCA) and (iii) regularizing or sparsity-inducing algorithms; e.g. EN, Ridge, and Dart (a boosted-tree method that randomly "drops out" trees and features). These techniques penalize overly complex models and help ensure that the forecasts generalize out of sample.

In Section 5.5 I examine the robustness of the training window approach. I compare the rolling and expanding window forecasts and find that the rolling window achieves superior results.

Performance assessment. I use a naïve benchmark for the performance assessment. Here, I use an historical average by taking the average PCA in-sample value to predict the OOS value. This naïve benchmark assumes that the PCA values move around their average value. I define the forecasting error of model m at time t for the ith PC as

$$\hat{e}_{i,t,m} = r_{i,t} - \hat{r}_{i,t,m}, \tag{5}$$

where $r_{i,t}$ denotes the realized return of the ith PC, and $\hat{r}_{i,t,m}$ denotes its forecast produced by model m. I follow Bali et al. (2023) using the standard out-of-sample definition of R^2:

$$R^2_{OOS,m} = 1 - \frac{\sum_{t=T_0}^{T} \hat{e}^2_{t,m}}{\sum_{t=T_0}^{T} \hat{e}^2_{t,\text{hist}}}, \tag{6}$$

where $\hat{e}_{t,m}$ is the forecasting error of model m and $\hat{e}_{t,\text{hist}}$ is the forecasting error of the historical average benchmark at t. I compute the measure for each PC and assess the performance separately. The measure compares the performance with a naïve benchmark of the historical average. The advantage of R^2_{OOS} compared to other measures like Mean Squared Error is that it is scale-invariant; it does not depend on absolute magnitude, but rather on relative magnitude compared to the benchmark. This is especially beneficial when comparing different PCs, which have different scales. For example, one PC may range from -10 to 10, while another may range from -5 to 5. Instead of the zero benchmark (as Bali et al. (2023)), I use

the historical average because PCs do not necessarily have an expected value of close to zero. If R^2_{OOS} is positive, the model performs better than its benchmark.

Similar to Bali et al. (2023), I use the Clark and West (2007) test to determine the statistical significance of the forecasts compared to the historical benchmark:

$$\text{CW} = \frac{\bar{c}}{\hat{\sigma}_c}, \tag{7}$$

where $\bar{c}$ denotes the time-series average and $\hat{\sigma}_c$ the Newey and West (1987) standard error of c_t. c_t is defined as the mean difference between the forecast error and its naïve benchmark

$$c_{t,m} = \frac{1}{n}\sum_{i=1}^{n}\left(e^2_{i,t,\text{hist}} - \hat{e}^2_{i,t,m}\right), \tag{8}$$

where n is the number of PCs that I predict. A positive value means that the model m performs better than the historical average.

Like Bali et al. (2023), I also compare the performance of different models. For example, I compare the performance of linear with non-linear models. To test if one model is signficantly better than another one, I use the Diebold and Mariano (1995) test. The test for comparing model A and B is defined as

$$\text{DM}^{(A,B)} = \frac{\bar{d}^{(A,B)}}{\hat{\sigma}_d^{(A,B)}}, \tag{9}$$

where $\bar{d}$ denotes the time-series average and $\hat{\sigma}_d$ the Newey and West (1987) standard error of d_t:

$$d_t = \frac{1}{n}\sum_{i=1}^{n}\left((\hat{e}^{(A)}_{i,t})^2 - (\hat{e}^{(B)}_{i,t})^2\right) \tag{10}$$

This approach enables the significance of forecasting performance to be evaluated and different forecasting methods to be compared. Note that for both evaluating individual model performance (using the CW test) and comparing models (using the DM test), I average forecast performance across PCs. While Bali et al. (2023) rely on a much larger cross-section, which likely yields more stable results, I be-

lieve this approach still provides meaningful insights. To prevent the risk of instable results, I also test the statistical significance of each PC across different models using the bootstrapping method of White (2000). The results can be found in the robustness checks in Section 5.5.

I further follow Bali et al. (2023) in assessing the linear similarity of forecasting results between model A and B by using the correlation between the forecasting errors. Contrary to to Bali et al. (2023), I do not use the cross-section to compute the correlation. I instead use the time-series correlation of error terms for each PC.

Interpretation. To draw meaningful conclusions about the predictors, and the performance that can be achieved, I assess the coefficients and the explanatory power of coefficients in each model. This is done differently for each model:

For the models Ridge, Lasso, and Elastic Net, I extract the coefficient values $\hat{\beta}$ and plot them. Coefficients that are zero or close to zero are penalized by the model and hence have less influence.

Plotting the raw coeffients cannot be done for the PLS regression. Here, I employ the variable importance in projection (VIP) metric following S. Wold, Sjöström, and Eriksson (2001) and Chong and Jun (2005). VIP gauges how strongly each predictor contributes to explaining the response through the PLS components: it scales the squared PLS weights by the proportion of Y-variance captured by each component. Following the rule-of-thumb of Chong and Jun (2005), a predictor with $\text{VIP}_j > 1$ is deemed influential.

Tree-based nonlinear models do not yield a single regression coefficient per predictor. Consequently, I evaluate predictors with the gain-based feature importance proposed by T. Chen and Guestrin (2016). Gain adds up the loss-reduction contributed by every split that uses a given feature (here the loss function is the squared error). I then normalize those gains so they sum to 1; each value shows that feature's share of the total loss-reduction achieved by the forest. Therefore, within each PC model, a larger feature importance implies that the predictor contributed more to lowering the training error. Neither VIP nor gain feature importance are causal coefficient values. They indicate how much the model relies on a variable to fit the data, but do not prove that changes in the variable cause changes in the response (Chong & Jun, 2005; T. Chen & Guestrin, 2016).

For simplicity, I do not extract the development of the coefficients over time. The coefficients and explanatory powers are based on the full sample size.

4.4 Economic performance

The relationship between the out-of-sample prediction performance of returns and its ability to produce economic value tends to be weak (Cenesizoglu & Timmermann, 2012). To assess the economic relevance of the PC forecasts, I conduct two steps. First, I follow Kagkadis et al. (2024) and convert predicted PC returns into factor returns. This is done by PC regression:

$$r_t = \delta^\top y_t + \varepsilon_t, \tag{11}$$

where $r_t \in \mathbb{R}^N$ is a column vector of factor returns (excess of risk free rate), $\delta \in \mathbb{R}^{7\times N}$ a coefficient matrix, $y_t \in \mathbb{R}^7$ is the column vector of PC returns, the latter including a one to include a constant. With the predicted one-month-ahead PCs $\hat{y}_{t+1}$ and estimated coefficients $\hat{\delta}$, I can then predict the factor returns $\hat{r}_{t+1} = \hat{\delta}^\top \hat{y}_{t+1}$.

Second, I construct portfolios by weighing the long-short option factor portfolios and evaluate risk-adjusted returns. I select the factor weights by solving the following optimization problem:

$$\max_{w_t} \frac{w_t^\top \hat{r}_t}{\sqrt{w_t^\top \hat{\Sigma}_t w_t}}, \qquad \text{s. t.} \quad w_t^\top 1_N = 1, |w_t| < 1_N, \tag{12}$$

where $w_t \in \mathbb{R}^N$ is the weight column vector and $\hat{\Sigma}_t \in \mathbb{R}^{N\times N}$ is defined as the historical diagonal matrix of variances of the past 60 factor returns. The estimation of covariances is tricky because I would have to estimate the entire covariance matrix with only 60 return observations per factor. I therefore set covariances between factor returns to zero in order to reduce noise (Ledoit & Wolf, 2022). Note that I restrict the weights to lay between -1 and 1 to avoid unrealisticly large bets that can arise in mean variance optimization. As a benchmark to the PC regression estimation of weights I construct a portfolio with equal weights $Nw_t = 1_N$.

5 Results

Table 4 exhibits key results. The first two PCs of option-related factors are predictable out-of-sample. The first PC captures idiosyncratic-volatility and low-quality risk, achieving an out-of-sample R^2 of 0.33 (linear ensemble) and 0.27 (nonlinear ensemble). The second PC represents equity-side frictions and limits-to-arbitrage,

with an R^2 of 0.15 for both ensembles. Thus, there is no advantage of using nonlinear models over linear ones to predict PCs of option returns. Generally, ML methods show the greatest improvement on the first PC, while for subsequent PCs, simple benchmarks like ARMA or even zero-return perform similarly.

Model	R^2_{OOS} **for PC1**	R^2_{OOS} **for PC2**	**Return**	**Sharpe Ratio**
Linear ensemble	0.33	0.15	0.21	1.95
Nonlinear ensemble	0.27	0.15	0.17	1.37
Equal-weight benchmark			0.04	2.71

Table 4: Out-of-sample predictability and economic performance of PC models. The table shows out-of-sample R^2 for the first two PCs of option-related factors, along with annualized returns and Sharpe Ratios (pre-transaction costs). Linear and nonlinear ensembles predict the first two PCs. The equal-weight benchmark represents the average performance of individual delta-hedged option factors. Sharpe Ratios are calculated using 3-month Treasury bills as the risk-free rate.

Both linear and nonlinear predictions generate economic gains. The linear ensemble returns 21% and the nonlinear ensemble returns 17%, both substantially outperforming the average delta-hedged option factor return. However, risk-adjusted performance is weaker. The equal-weighted benchmark achieves a Sharpe Ratio of 2.71 versus 1.95 (linear) and 1.37 (nonlinear), despite its lower 4% return.

In the following, I explain these results in four parts. First, I describe the main PCs of factor returns and highlight which types of factors are more predictable than others. Second, I evaluate the forecasting performance of the ML models employed. Third, I examine whether these forecasts generate economic value by constructing portfolios based on their out-of-sample performance. Lastly, I conduct a series of robustness checks and discuss the results.

5.1 Dimensionality reduction

My first six PCs capture linear patterns in the factor return data. The first six PCs explain about 69% of the total linear variance, based on the share of the sum of the first six eigenvalues relative to the sum of all eigenvalues. Hence, option factors can be reduced to a couple of components - similar to equity factors.[9] Table 5 shows the factor loadings for each variable. Factor loadings vary among different

[9]Haddad et al. (2020) find that the first five PCs explain approximately 60% of linear variation in equity factor returns. This finding challenges the "factor zoo", because many factors can be broken down into fewer components.

factors. I characterize the PCs using loadings above or below 0.5 or -0.5.[10]

Factor	PC1	PC2	PC3	PC4	PC5	PC6
log_price	0.63	0.59	-0.12	0.04	0.16	-0.03
netis_at	0.19	-0.26	-0.43	-0.10	-0.43	0.35
ebit_sale	0.67	-0.19	-0.24	0.29	-0.14	-0.10
ope_be	0.77	-0.22	-0.27	0.12	-0.11	-0.01
ocfq_saleq_std	0.76	-0.09	-0.22	0.09	-0.12	0.02
cash_at	0.22	-0.12	-0.47	-0.04	-0.59	0.06
disp	0.72	0.13	0.08	-0.14	0.11	0.23
zscore	-0.18	0.61	-0.07	-0.35	0.05	0.42
issue_5y	0.63	-0.21	-0.22	-0.03	0.04	0.08
issue_1y	-0.13	0.26	-0.29	-0.52	-0.02	0.44
rsi	0.46	0.20	0.02	-0.29	-0.11	-0.03
amihud	0.43	0.65	-0.33	0.22	0.09	-0.28
ivrv	0.16	0.66	0.37	0.30	-0.29	0.13
ivol	0.86	-0.28	0.02	-0.03	0.01	-0.04
tskew	0.12	-0.22	-0.20	0.43	0.47	0.42
iskew	0.19	-0.15	-0.08	0.51	0.43	0.51
defrisk	0.56	0.54	0.25	-0.30	0.19	0.16
max10	0.80	-0.33	0.03	-0.20	0.03	-0.01
ac	0.37	0.01	0.26	-0.06	-0.03	0.07
ivterm	0.59	0.22	0.52	0.15	-0.05	-0.13
sysvol	-0.77	0.42	-0.10	0.20	-0.17	0.08
hvol	0.89	-0.27	0.04	-0.10	0.05	0.01
optspread	-0.09	-0.53	0.55	0.11	-0.24	0.25
vov	0.19	0.10	0.32	0.15	-0.38	0.34
jr	-0.37	0.30	-0.12	0.39	-0.34	0.12
vr	0.83	0.13	0.22	0.02	-0.15	0.03
hc	0.62	0.55	-0.30	0.18	0.04	-0.20
embedlev	0.87	0.09	0.20	0.09	-0.16	0.04

Table 5: Loadings of the first six PCs on each factor (rounded to two decimal places).

1. *Idiosyncratic-volatility & low-quality risk*: The first PC is positively correlated with most factors. Therefore, I observe both option- and stock-based features. It includes volatility-based risk factors, such as historical firm volatility (hvol) and idiosyncratic volatility (ivol). It also includes firm-level quality characteristics, such as operating profitability (ope_be) and profit margin (ebit_sale). Note that stocks with high idiosyncratic volatility tend to have lower systematic volatility (sysvol). PC1 generally captures firms that are cheap, unprofitable, cash-flow volatile, and highly leveraged, whose options display very high idiosyncratic (rather than systematic) volatility.

[10] Alternatively, one could pursue dimensionality reduction through factor analysis using *orthogonal* or *oblique* rotations. These methods offer more interpretable results than factor loadings of PCA (Fabrigar & Wegener, 2011).

2. *Equity-side frictions & limits-to-arbitrage*: The second PC has a positive loading with factors equity illiquidity (amihud), Altman Z-score (zscore), and default risk (defrisk) that cover equity side friction. The loadings for option-side market frictions, such as delta-hedging costs (hc), option illiquidity (optspread), and implied-minus-realised volatility (ivrv) include option-based volatility measures. Option bid-ask spreads load negatively, so the factor is a wedge between stock-side and option-side frictions. In total, the second PC covers hard-to-trade stocks and relatively tighter option markets.

3. *Option illiquidity & volatility term structure*: The third PC only has two strong factor loadings. As option illiquidity (optspread) and implied ATM volatility term structure (ivterm) are positive loadings, this PC emphasizes the frictions in the options markets. Together with the second PC, the third PC covers the trading-friction surface.

4. *Idiosyncratic skew & recent issuance*: The fourth PC includes stocks that issued equity in the past year (issue_1y) as well as stocks with high idiosyncratic skewness (iskew). The PC identifies "lottery" names (large skew) with little recent issuance.

5. *Financial slack*: Orthogonal to the big volatility/ liquidity stories, the fifth PC captures the cash position of the underlying stocks (cash_at).

6. *Residual skew*: The sixth PC captures residual skewness of the underlying stock (tskew), complementing the fourth PC.

In summary, the first PC signals quality-volatility, the second and third PC captures market-frictions, while the remaining PCs explain tail-risk and financing nuances. Figure 4 shows a biplot of the first two PCs. Together, the two PCs explain about 45% of the linear variation. The graph confirms the factor loadings described earlier.

5.2 Model forecasting performance

Table 6 shows the R^2_{OOS} for each model.[11] Several observations can be made: Firstly, forecasting performance varies across PCs. It appears that the ML mod-

[11]Graphs and tables use the following short names for models; *hist*: historical mean return over the training period, *arima*: ARMA model, *lasso*: Lasso, *ridge*: Ridge, *en*: Elastic Net, *pls*: Partial Least Squares, *linear_ens*: equally weighted average of linear models, *rand_forest*: Random Forest, *XGBoost*: XGBoost, *Dart*: Dart, *nonlinear_ens*: equally weighted average of nonlinear models, *zero_return*: constant zero PC forecast.

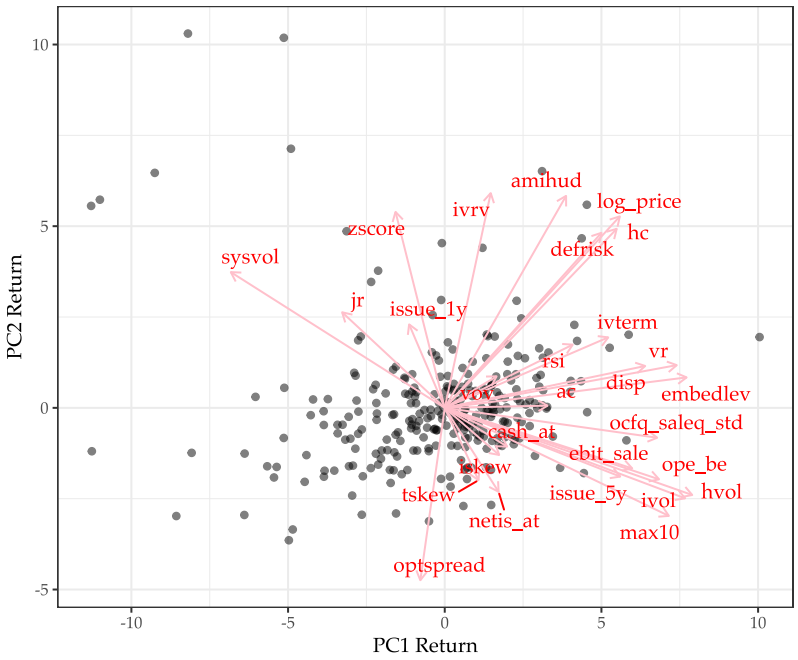


Figure 4: This PC1 and PC2 biplot puts into context the factor loadings of the factors for the two most important PCs. The scatterdots are the PC returns. The arrows illustrate the factor loadings of the factors. The gray points are spread wider horizontally than vertically, indicating that PC1 contributes more to return variance than PC2. This mirrors the eigenvariance ranking. The visual representation of the factor loadings corroborates earlier observations. Many arrows point to the left, most of which are volatility or quality factors. The only prominent factor pointing to the left is *sysvol*, which, as discussed earlier, is negatively related to the other factors in PC1. There are also upward arrows indicating equity-side frictions, as well as an upward-pointing arrow indicating option-based frictions (bid-ask spread). Note that factors such as *cash_at* and *tskew* have very short arrows and are therefore mostly captured by PCs three through six. Find more visualizations in Appendix D.

els can predict the first and second PCs more accurately than the others. Linear and nonlinear ensemble models achieve an R^2_{OOS} of up to 33% and 27% for the first, respectively, and 15% for the second PC. The first PC embodies the strongest common signal in the factor space: idiosyncratic volatility combined with low-quality fundamentals. The second PC covers equity-side frictions and limits to arbitrage. My predictor set may be especially suited to reliably forecast return patterns related to these characteristics. Additionally, lower-order PCs represent progressively "leftover" variation. Smaller eigenvalues imply lower signal-to-noise ratios and a heavier mix of measurement error and transitory shocks. Consequently, it is more difficult for all algorithms to extract stable patterns from those components (Jolliffe, 2002). Secondly, there are differences between the ML models. For example, PLS performs worse than the naïve historical benchmark for five out of

six PCs, whereas Ridge achieves a positive R^2_{OOS} for the first four PCs. Thirdly, benchmarks such as the zero-return and the ARMA model perform poorly for the first PC. Nevertheless, they achieve similar R^2_{OOS} values to the ML models for the second PC, indicating no advantage in using ML methods to capture equity-side frictions and limits to arbitrage.

Model	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**
arima	0.03	0.17	-0.09	-0.00	-0.02	0.03
zero_return	0.04	0.17	-0.08	0.02	0.01	-0.02
lasso	0.31	0.04	-0.11	0.00	-0.07	-0.25
ridge	0.17	0.19	0.07	0.04	0.01	-0.01
en	0.34	0.04	-0.03	-0.05	-0.11	-0.23
pls	0.09	-0.05	-0.26	-0.21	-0.34	-0.31
linear_ens	0.33	0.15	0.02	0.05	0.03	-0.12
rand_forest	0.19	0.11	-0.06	-0.13	-0.12	-0.18
XGBoost	0.26	0.10	-0.12	-0.06	-0.15	-0.20
Dart	0.28	0.16	-0.04	-0.07	-0.08	-0.09
nonlinear_ens	0.27	0.15	-0.02	-0.04	-0.07	-0.11

Table 6: Out-of-sample R^2 for each model. Positive numbers indicate superior performance compared to the naïve benchmark of the historical average. It is evident that all models perform better for the first PC, and almost all models perform better for the second PC. The performance decreases with the PC rank: after the second PC, models do not or only slightly outperform the benchmark. The ensemble models have the highest R^2, while the linear ensemble performs slightly better than the nonlinear one.

To check the statistical significance of the forecast compared to the naïve historical average, I compute the Clark and West (2007) test statistic for each model. The results are shown in Table 7. The models produce different outcomes. All linear models except for PLS achieve statistically improved forecasts compared to the zero benchmark at a 1% significance level. Of these, the Ridge performs best, with a test statistic of 3.79. All nonlinear models achieve superior forecasts (again, at a 1% significance level). For both linear and nonlinear models, the equally weighted forecast ensembles achieve the largest test statistics, confirming the merits of using ensemble forecasts as outlined by Timmermann (2018).

Table 8 uses the Diebold and Mariano (1995) test to compare every pair of forecasting models. ML approaches generally outperform the naïve benchmarks: most ML–benchmark entries are positive and significant at the 5% level (yet, PLS, XGBoost, and Random Forest do not achieve to outperform benchmarks). In contrast, within the ML camp, the picture is more nuanced. With the exception of two clear under-performers (the PLS on the linear side and Random Forest on the non-linear

Model	Mean	SE	CW
zero_return	0.10	0.05	2.03
lasso	0.29	0.10	**2.95**
ridge	0.28	0.07	**3.79**
en	0.33	0.10	**3.37**
pls	-0.09	0.12	-0.77
linear_ens	0.42	0.09	**4.57**
arima	0.08	0.06	1.23
nonlinear_ens	0.32	0.08	**4.18**
XGBoost	0.24	0.10	**2.46**
Dart	0.32	0.08	**3.83**
rand_forest	0.17	0.06	**2.83**

Table 7: Clark and West (2007) test to assess significance of outperformance. The one-sided standard-normal critical values are 1.65 (5%), 1.96 (2.5%), and 2.33 (1%). Test statistics larger than the 1% critical value are highlighted in bold. Except for the PLS and ARMA models, all models have significantly better performance than the historical average at the 5% confidence level. Even the zero-return benchmark achieves this outperformance, mainly driven by better predictions. Both ensemble methods achieve outperformance at the 1% significance level.

side), linear and nonlinear methods yield statistically insignificantly different accuracy. Lasso, Ridge, EN, XGBoost, and Dart in particular are never significantly outperformed by a single model. The ensemble forecasts do improve upon a number of the individual models they pool, even yielding significantly superior results compared to some individual models that they are composed of.

Model	lasso	ridge	en	pls	linear_ens	arima	nonlinear_ens	XGBoost	Dart	rand_forest
zero_return	1.96	2.54	2.38	-1.68	**3.65**	-0.32	**2.94**	1.42	**2.66**	1.26
lasso		-0.13	0.89	-4.22	2.26	-1.86	0.33	-0.55	0.30	-1.36
ridge			0.60	**-3.22**	**2.66**	-2.23	0.76	-0.56	0.65	-2.19
en				**-4.37**	2.55	-2.13	-0.04	-1.06	-0.07	-1.98
pls					**5.73**	1.50	**3.49**	**2.62**	**3.50**	2.18
linear_ens						**-3.43**	-1.72	**-2.68**	-1.66	**-3.86**
arima							**2.83**	1.51	**2.81**	1.27
nonlinear_ens								-2.45	-0.13	**-4.43**
XGBoost									1.92	-1.14
Dart										**-3.34**

Table 8: Diebold and Mariano (1995) test to compare model performance. The two-sided standard-normal critical values are 1.65 (10%), 1.96 (5%), and 2.58 (1%). Test statistics larger and smaller than the 1% critical values are highlighted in bold. Positive test statistics indicate that the column model performs better than the row model, and vice versa. The linear ensemble outperforms all models it is composed of, including EN, Ridge, Lasso, and PLS. It also performs better than the nonlinear models XGBoost and Random Forest. While the nonlinear ensemble is significantly superior to some models (such as Random Forest and XGBoost), it performs significantly worse at only the 10% confidence level.

Nonlinear models do not appear to outperform linear models in the prediction of

PCs. There are several possible reasons for this: (i) The time-series behaviour of PCs may not embed as many nonlinear patterns as expected. (ii) The sample size may be too limited, or the data may be too noisy, resulting in overfitted models. (iii) The hyperparameters may be tuned too rarely, or the space may be defined incorrectly.

In Table 9, I plot how correlated forecast errors are between models. I find that forecast errors ar strongly and positively correlated across all models. Generally, correlation is higher between linear or between nonlinear models. For example, the correlation of Dart with Lasso, Ridge, and Elastic Net are 0.88, 0.93, and 0.89, respectively. Conversely, the correlation of Dart with XGBoost and Random Forest is 0.95 and 0.94, respectively. Therefore, creating an ensemble of linear and nonlinear models could be beneficial, since uncorrelated forecasting methods can enhance overall prediction performance (Timmermann, 2018).

	real_return	lasso	ridge	en	pls	linear_ens	hist	arima	nonlinear_ens	XGBoost	Dart	rand_forest
real_return		0.82	0.94	0.82	0.66	0.86	0.96	0.91	0.94	0.88	0.89	0.99
lasso			0.89	0.97	0.79	0.96	0.83	0.82	0.89	0.87	0.88	0.86
ridge				0.90	0.78	0.94	0.96	0.93	0.95	0.92	0.93	0.96
en					0.80	0.97	0.83	0.82	0.89	0.88	0.89	0.86
pls						0.90	0.67	0.71	0.75	0.74	0.75	0.71
linear_ens							0.87	0.87	0.92	0.90	0.91	0.89
hist								0.91	0.93	0.88	0.89	0.96
arima									0.90	0.86	0.88	0.92
nonlinear_ens										0.98	0.99	0.97
XGBoost											0.97	0.93
Dart												0.94
rand_forest												

Table 9: Correlation of model forecasts across all PCs. Almost all forecasts are highly correlated with each other. Linear models (Lasso, Ridge, Elastic Net, and their ensemble) as well as nonlinear models (XGBoost, Dart, Random Forest, and their ensemble) form tightly linked groups with correlations above 0.9. PLS stands out with notably lower correlations, especially with the historical average (0.67) and ARIMA (0.71). As forecasts from PLS are less correlated, including PLS in model ensembles could increase forecasting performance. (Timmermann, 2018)

To check how forecasting performance evolves over time, Figure D4 plots the smoothed squared error. The figure shows that the forecasting errors develop similarly over time for each model and PC, but differ in terms of absolute values. The first two PCs show higher errors in the periods 2000 to 2005, 2007 to 2010, and after 2017. The high errors during the first two periods could be due to heavy financial turbulence during the dotcom crisis and the global financial crisis (GFC).

As most ML models perform similarly, and ensemble methods achieve the best results (see Table 8), I continue with the two ensemble models: *linear_ens* and *nonlinear_ens*. Focusing on these two methods simplifies the comparison between linear and non-linear estimates.

5.3 Feature importance

Figures D6 to D9 illustrate how individual features contribute to the prediction of linear models. Although multiple features contribute more than others across PCs, there is significant variation across models and PCs. Significant features do not necessarily have to correspond with factor loadings. For instance, the fifth PC - with the main factor loading being *cash_at* - is mainly predicted by leverage (*netis_at*), profitability (*ebit_sale*), and characteristic spread (*spread*2) across all models. Strikingly, factor momentum (especially *mom*3 and *mom*4) are relevant coefficients for most PCs and models, mainly for linear models, confirming results of Käfer, Mörke, and Wiest (2025). In contrast, for PLS, feature importance - measured by VIP - indicates strong relevance of PC momentum, being the most relevant feature across all PCs.

Figures D10 to D12 show the importance of features in reducing the squared error. Generally, the overall option market return (*ew_ret*) is among the most important factors across models for the first PC. Most features that contribute substantially to error reduction arenot option-specific. An exception is seen in the Random Forest model, where *embedlev* and *issue_1y* play the largest role in predicting PC6. Feature importance differs notably between Random Forest and the other models, but is similar between XGBoost and Dart.

Overall, the strong variation in feature importance makes it difficult to identify consistent key indicators. All types of predictors show relevance for specific PCs and models. This includes factor weights, characteristics such as momentum, non-tradable factors, and general market indicators like investor sentiment. The results suggest that using a wide range of predictors is beneficial, especially when applying models with regularisation to prevent overfitting.

5.4 Economic performance

Table 10 presents the returns of each model portfolio. The returns of ML ensemble models outperform the benchmarks, with the historical average, ARMA, and equal-weighted portfolio having the lowest returns. The linear ensemble achieves the highest annualised return of 21%. As discussed in Section 5.2, forecasting performance does not directly relate to higher returns. Still, we observe that ensemble models - that have the highest predictive performance - also achieve the highest returns.

Model	Return	Total return	SD	Sortino Ratio	Hit Ratio	Sharpe Ratio	SR CI low	SR CI high
arima	0.11	2.25	0.10	2.03	0.67	1.10	0.70	1.56
equal_weight	0.04	1.11	0.02	3.96	0.85	2.71	2.15	3.40
hist	0.07	1.67	0.09	1.44	0.66	0.81	0.38	1.27
linear_ens	0.21	4.23	0.11	5.04	0.78	1.95	1.58	2.36
nonlinear_ens	0.17	3.44	0.12	3.13	0.71	1.37	1.07	1.81

Table 10: Model performance overview. *Return*: annualized excess return; *Total return*: the return over the entire period; *SD*: annualized standard deviation of returns; *Sortino Ratio*: downside-risk adjusted return ($\bar{r}/SD(r^-)$, where $SD(r^-)$ is the standard deviation of negative returns); *Hit Ratio*: fraction of periods with positive returns; *Sharpe Ratio*: risk adjusted returns ($r - \bar{r}_{rf}/SD(r)$); *SR CI low* and *SR CI high*: 95% confidence interval. Confidence bands are created by bootstrapping, following Ledoit and Wolf (2008): I take 1000 bootstrap samples from the return series (with replacement) and from these simulations retrieve the confidence bands of the Sharpe Ratio. For the risk-free rate r_{rf}, I use the 3-Month Treasury Bill Secondary Market Rate (Discount Basis).

Adjusting the annual returns for risk tells a different story. There, the equally weighted portfolio outperforms all the other models. Its Sharpe ratio exceeds 2.7 and its confidence interval only overlaps with that of the linear ensemble. The equal-weighted portfolio achieves this through its very low variance. It also excels in the Hit Ratio, indicating more consistent positive profits. However, the linear ensemble has the highest Sortino ratio, indicating that negative returns are less volatile for the linear ensemble. Hence, neither linear nor nonlinear estimation methods outperform the equal-weighted benchmark across different measures. The equal-weighted benchmark generally achieves the best risk-adjusted performance, with the linear ensemble performing similarly well. Compared to the nonlinear ensemble, the linear one performs better. Figure 5 illustrates the Sharpe Ratios visually.

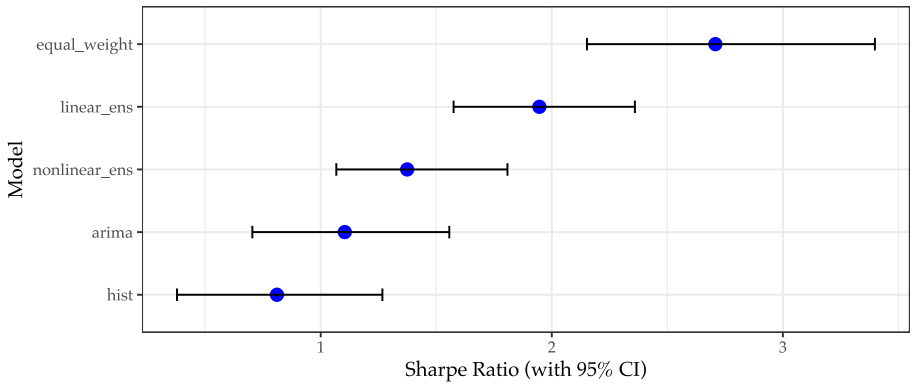


Figure 5: Annualized Sharpe Ratios with 95% confidence bands. Confidence bands are created by bootstrapping, following Ledoit and Wolf (2008): I take 1000 bootstrap samples from the return series (with replacement) and from these simulations retrieve the confidence bands.

Comparing raw and risk-adjusted returns with those of individual factors shows that combining factors is beneficial. The positive forecasting performance of the models allows them to achieve returns that are more than double the average of individual factors. Yet, this is mainly achieved by taking on more risk, as the Sharpe Ratio remains similar to that of individual factor portfolios. The benefits of diversifying factor risks become apparent when examining the equally weighted factor portfolio, which, by definition, has an average factor return but diversifies factor risk and achieves a high Sharpe Ratio.

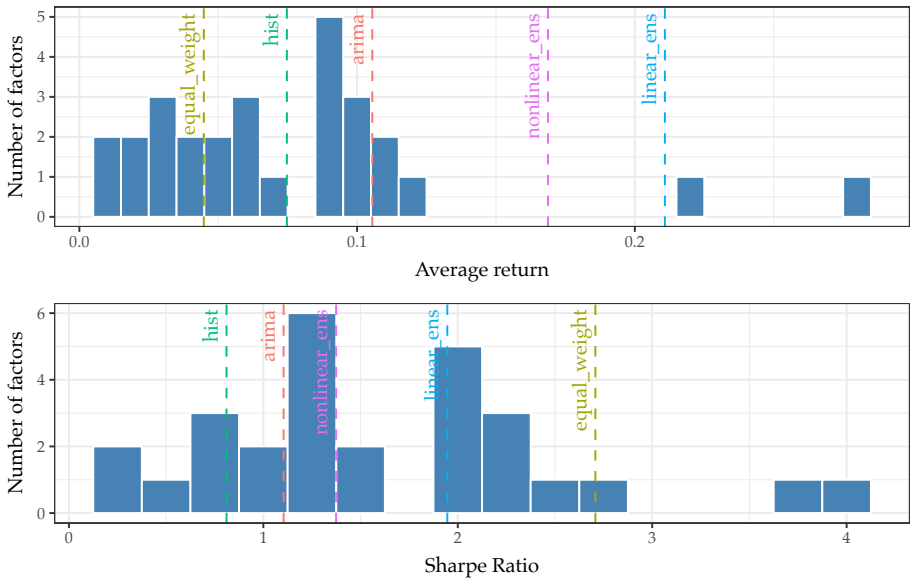


Figure 6: Annualized model returns and Sharpe Ratios compared to individual factor portfolios. While the linear and nonlinear ensembles achieve higher returns than the average factor return, they only achieve around average factor return Sharpe Ratios, while the equal-weights portfolio generally achieves a higher ratio.

One potential reason models perform well in raw returns but not in risk-adjusted returns is the construction of the covariance matrix. Several problems arise in covariance estimation: First, I only use the historical variances of factor returns. This method reacts slowly to new patterns and weighs all returns equally.[12] Second, I exclude all covariations of factor returns. Since factor returns comove, as

[12]To address this issue, I incorporate a simple volatility estimate using an the RiskMetrics approach (Pafka & Kondor, 2001). I forecast the variance using $\sigma^2_{i,t} = (1-\lambda)\sum_{\tau=1}^{T}\lambda^{\tau}(y_{i,t-\tau} - \mu_{i,t-\tau})^2$, where $\sigma^2_{i,t}$ is the return variance and $\mu_{i,t}$ the expontially weighted return mean of factor i at t. I set $\lambda = 0.94, T = 60$. The original RiskMetrics approach does not include the mean value. Because option factor returns achieve high positive return, I adapt the framework to incorporate them. However, this approach does not significantly increase the Sharpe Ratio of the models. Find more details in Table D7.

demonstrated in Section 5.1, this eliminates a significant amount of information, particularly that relevant to diversification benefits. Due to limited data availability (e.g., monthly data only and a 60 months time horizon), there are limitations to volatility and covariance modeling. This makes multivariate GARCH models not a viable alternative, as they require the estimation of many coefficients and parameters.[13] To solve this issue, I test the linear shrinkage to identity matrix of Ledoit and Wolf (2022). However, as the covariance estimates are extremely noisy, the shrinkage proposes the usage of the diagonal matrix of average variances over all factors. Find the Results in D8.[14] Also, estimating the covariance matrix with an expanding window does not significantly increase the Sharpe Ratios. Alternatively, increasing the sample size T by using higher-frequency data (e.g., daily factor returns) or estimating the covariance through other shrinkage methods, like nonlinear shrinkage, could improve prediction accuracy and produce more stable covariance outcomes (Ledoit & Wolf, 2022).

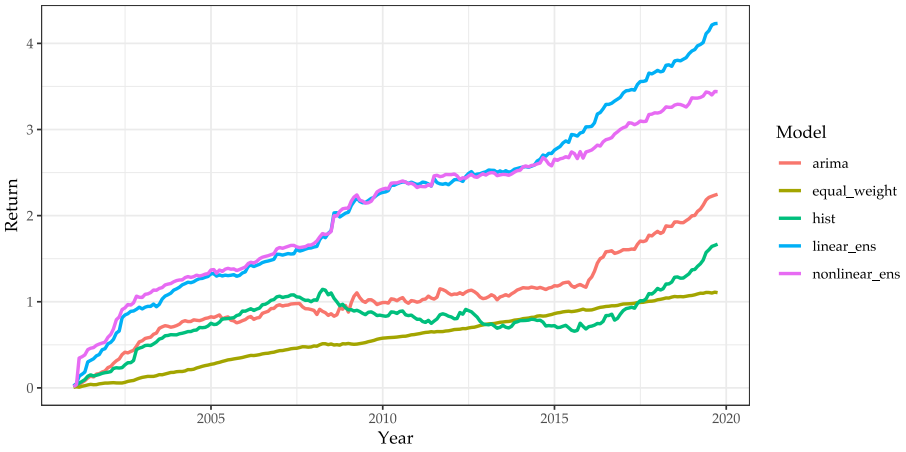


Figure 7: Cumulative monthly returns of optimal portfolio allocation per model.

Figure 7 shows the cumulative returns of each strategy. It illustrates how the returns evolve over time, highlighting two significant observations. First, the ensemble ML models exhibit strong comovement, while the benchmark models and the equal-weighted portfolio appear to follow distinct patterns. For example, while all

[13] For instance, the CCC-GARCH model necessitates estimating the correlation matrix and univariate variance processes. The original CCC-GARCH includes a total of $N(N+5)/2$ parameters. With 28 factors, this equals to 462 parameters (Bauwens, Laurent, & Rombouts, 2006).

[14] Using a single variance across all factors increases the return even further. For example, the linear ensemble achieves an annual return of 26%. This is because the optimizer focuses solely on optimizing returns. All possible portfolios have the same variance by design, regardless of the weight chosen.

ensemble ML models experience large returns during the GFC, their benchmarks have negative returns. This can be explained by the fact that all of the models use predicted PCs in the same regression to estimate returns and base their weights on that information. The ARMA benchmark and historical average benchmarks likely do not comove as much because they fail to accurately predict PCs. Second, there are periods in which the models perform better and others in which they perform worse. Three periods of strong growth for most ML models are: (i) the early 2000s, (ii) the GFC, and (iii) after 2015. In contrast, the period between 2010 and 2015 shows close to zero returns for most models. Since two estimations are in play - first, the estimation of PCs, and then, the estimation of returns - it is difficult to determine the reason for these performance differences. I discuss structural breaks and the role of the GFC more extensively in the robustness checks in Section 5.5.

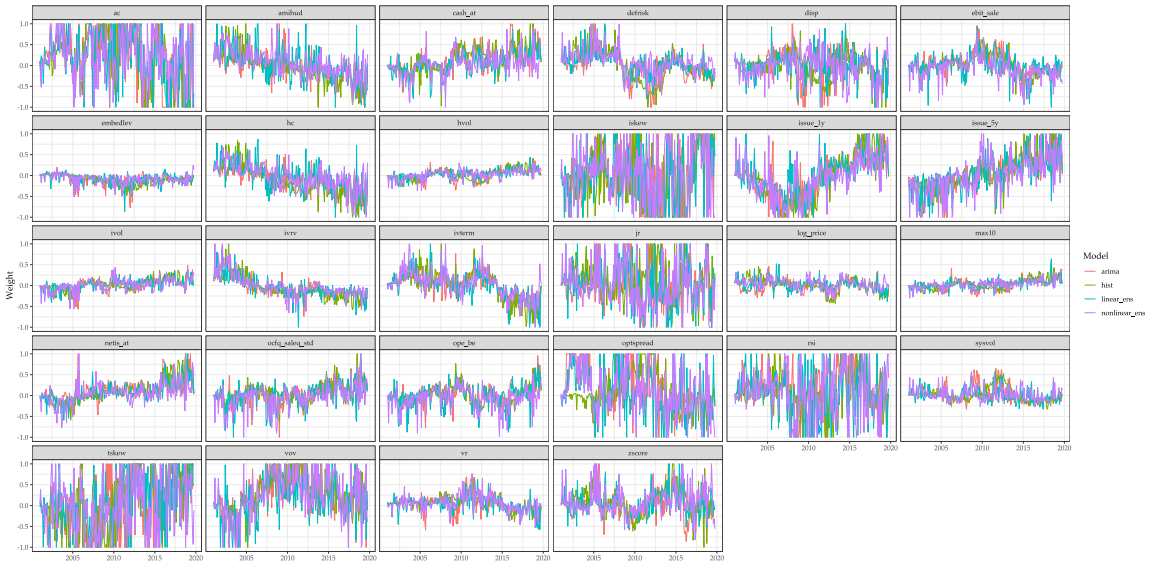


Figure 8: Optimal portfolio weights per model and factor. Factor weights are restricted between 1 and -1. Some factor weights change over time, others do not. For example, *embedlev* stays close to zero throughout. *amihud* starts positive and trends negative. Other factor weights show extreme swings. *iskew* oscillates sharply between -1 and 1, while *hvol* remains relatively stable. We can make multiple observations in relation to the GFC and the monetary normalization of 2015. The GFC triggers lasting shifts. For instance, *defrisk* turns negative as default sensitivity increased. In contrast, *ebit_sale* gains weight as profit margins signals resilience (Zhan et al., 2022). Similarly, *vr* spikes during the crisis; this could be due to increased volatility hedging demand. Post-GFC, *disp* oscillates more amid persistent uncertainty. Post-2015 monetary normalization impacts other factors. Some examples are: *ivterm* declines, as the volatility term structure flattens, *zscore* oscillates strongern, and *netis_at* turns positive ("Risk Management in Central Banks in the Context of Monetary Policy Normalization", 2019). Although factor risk premia for stocks during changes in monetary policy or financial crises are well researched, option factors during these periods have yet to be studied.[15]

Figure 8 shows how factor weights change over time, especially during major changes in the financial markets, such as the GFC. Haddad et al. (2020) highlight that equity factors exhibit heterogeneous time-variation patterns. Additionally, factor loadings themselves are state-dependent, meaning they depend on economic conditions. Strong variation in optimal factor weights for option factors could indicate that the observations made by Haddad et al. (2020) also apply to option factors.

Figure 9 shows the volatility of factor timing portfolios, which increases during times of financial stress, such as the GFC. These observations suggest that factor timing portfolios are exposed to volatility during turbulent periods, yet still generate positive returns.

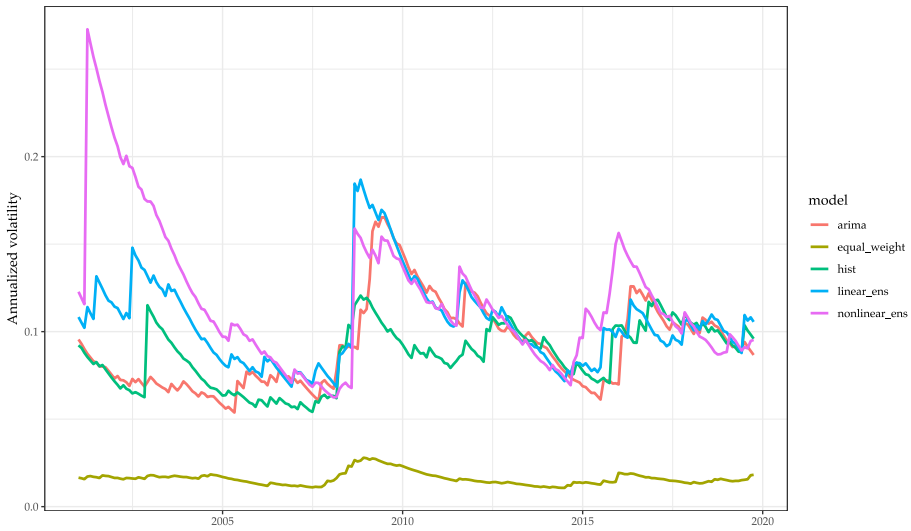


Figure 9: Annualized volatility based on the exponential weighted moving average (EWMA) method: $\sigma_t^2 = \lambda\sigma_{t-1}^2 + (1-\lambda)(r_{t-1} - \mu_{t-1})^2$, where $\mu_t = \lambda\mu_{t-1} + (1-\lambda)r_{t-1}$ and $\lambda = 0.94$. Three volatility spikes are visible: one at the beginning of the 2000s, one during the GFC, and one after 2015 during the normalization of monetary policy. Notably, returns tend to increase during these periods. Although the equally weighted portfolio exhibits significantly lower overall volatility than every other portfolio, it too experiences volatility spikes during the GFC and post-2015.

[15]For instance, Gourio (2013) links default risk premia to disaster sensitivity. The author builds a model where credit spreads are high and volatile due to risk premia. These premia reflect the chance of rare economic disasters. Firms are exposed to such disasters through default risk. The effect on option risk premia remains unstudied.

5.5 Robustness checks

My finding that there are predictive patterns in option factors that translate into non-risk-adjusted economic gains is based on my specific model setting. Before generalizing the findings, I must conduct reality checks on several different specifications, such as data selection, estimation methods, and transaction costs. Below, I discuss the decisions I made to verify the robustness of my results.

Factor return computation. In contrast to Kagkadis et al. (2024), Haddad et al. (2020) estimate optimal factor weights directly in the PC space. This avoids a second regression step to obtain factor weights, potentially reducing noise. Since PCs are linear combinations of factor returns, I can first find the optimal PC weights w_{PC}, and then recover the corresponding factor weights as $w_R = Q_R w_{PC}$, where Q_R contains the PC loadings. Given that the factor returns are computed as $r = PC \cdot Q_R^\top$, this transformation is valid. To ensure the factor weights satisfy the same constraints as in Equation 12, I must impose $w_{PC}^\top Q_R^\top 1_N = 1$ and $-1 \leq Q_R w_{PC} \leq 1$.

Model	Return	Total return	SD	Sortino Ratio	Hit Ratio	Sharpe Ratio	SR CI low	SR CI high
arima	0.13	2.78	0.07	4.51	0.75	2.03	1.61	2.48
equal_weight	0.04	1.11	0.02	3.96	0.85	2.71	2.15	3.40
hist	0.14	2.87	0.07	3.21	0.78	2.04	1.59	2.59
linear_ens	0.20	3.96	0.08	6.77	0.81	2.43	2.03	2.96
nonlinear_ens	0.15	3.09	0.07	3.70	0.81	2.10	1.66	2.65

Table 11: Model performance overview, optimizing PC weights directly. (Haddad et al., 2020) *Return*: annualized excess return; *Total return*: the return over the entire period; *SD*: annualized standard deviation of returns; *Sortino Ratio*: downside-risk adjusted return ($\bar{r}/SD(r^-)$, where $SD(r^-)$ is the standard deviation of negative returns); *Hit Ratio*: fraction of periods with positive returns; *Sharpe Ratio*: risk adjusted returns ($r - \bar{r}_{rf}/SD(r)$); *SR CI low* and *SR CI high*: 95% confidence interval. Confidence bands are created by bootstrapping, following Ledoit and Wolf (2008): I take 1000 bootstrap samples from the return series (with replacement) and from these simulations retrieve the confidence bands of the Sharpe Ratio. For the risk-free rate r_{rf}, I use the 3-Month Treasury Bill Secondary Market Rate (Discount Basis).

The results are shown in Table 11. While this direct approach does not increase returns for most models, Sharpe Ratios are significantly higher. This observation is likely due to the fact that estimating the covariance of six PCs involves less noise than estimating it for 28 factor returns. With only six weights to optimize, I use the PCs' historical covariance matrix over the past five years. Note that the naïve benchmark models, from which I now benefit from more accurate covariance estimation, also achieve high Sharpe Ratios greater than two while still lagging behind ensemble ML methods. These Sharpe Ratios support the idea that covariance estimation introduces too much noise. Generally, using the PCs directly, as in Haddad

et al. (2020), increases returns.

Transaction costs. Ofek, Richardson, and Whitelaw (2004) show that transaction costs significantly reduce the returns of option portfolios. Similarly, Detzel, Novy-Marx, and Velikov (2023) note that ignoring transaction costs lead to favoring high-cost factors. Thus, accounting for transaction costs is crucial for a proper assessment of factor timing returns and the overall economic impact of factor timing strategies.

Since the dataset does not include actual transaction costs, I use an option illiquidity proxy. I construct this proxy by calculating the option half-spread in dollar terms and scaling the value by 20.3%. The resulting value is then multiplied by two to account for both the opening and closing of a position. The construction of the transaction cost proxy can be expressed as: TC Proxy $= 2 \cdot \text{optspread} \cdot \text{mid} \cdot \text{scale}/\text{denominator}$. To construct factor returns net of transaction costs, I deduct the average transaction cost proxy of the options within the long and short deciles: $\text{Factor Return}_{\text{long}} - \text{Factor Return}_{\text{short}} - (\text{Avg. TC Proxy}_{\text{long}} + \text{Avg. TC Proxy}_{\text{short}})$ (Muravyev & Pearson, 2020; Heston, Jones, Khorram, Li, & Mo, 2023; Käfer, Moerke, et al., 2025).

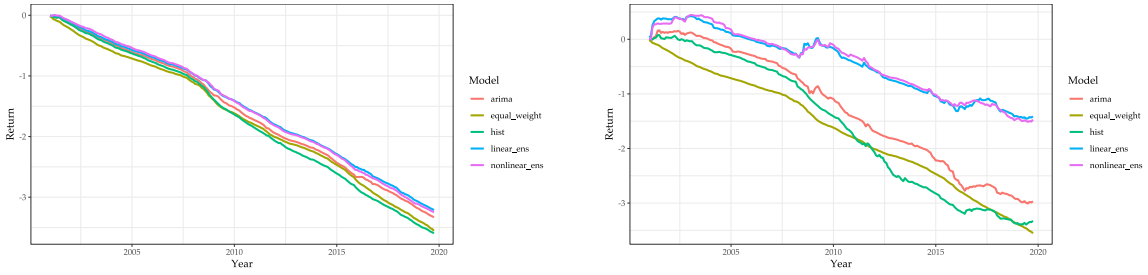
Model	PC1	PC2	PC3	PC4	PC5	PC6
arima	0.21	0.01	0.06	-0.07	0.01	-0.03
zero_return	0.12	0.15	-0.01	-0.06	0.06	-0.04
lasso	0.45	0.06	0.01	-0.12	-0.03	-0.10
ridge	0.34	0.21	0.08	-0.01	0.11	0.04
en	0.47	0.05	0.04	-0.03	0.01	-0.15
pls	0.32	0.12	-0.40	-0.50	-0.01	-0.14
linear_ens	0.47	0.21	0.05	-0.03	0.10	-0.02
rand_forest	0.27	0.17	-0.08	-0.15	-0.05	-0.20
XGBoost	0.44	0.29	0.04	-0.11	-0.00	-0.05
Dart	0.39	0.28	0.06	-0.05	0.01	-0.10
nonlinear_ens	0.40	0.28	0.05	-0.05	0.02	-0.06

Table 12: Out-of-sample R^2 for each model for predicting PCs of factor returns, net of transaction costs.

Option factor returns net of transaction costs increases the predictability of factor returns. Table 12 shows that linear and nonlinear ensemble estimations achieve out-of-sample R^2 of 47% and 40%, respectively. The prediction quality hence is higher than without accounting for transaction costs. However, this improvement does not translate into positive returns.

There are two approaches to constructing optimal factor timing portfolios: either

include all factors, or include only those predicted to be positive (following Goyal and Saretto (2022)). To prevent look-ahead bias, I check only whether *predicted* future returns are positive. Cumulative returns are shown in Figure 10. Although returns are slightly less negative for the strategy including all factors, neither strategy achieves positive returns over the sample period. Therefore, transaction cost mitigation strategies should be applied for predictability to translate into economic gains.



(a) Excluding negative returns **(b)** Including all returns

Figure 10: Cumulative returns under different transaction cost assumptions. The left panel excludes negative returns; the right panel includes all returns. Both panels assume 20.3% of the option half spread as transaction costs, paid two times for opening and closing of position. (Muravyev & Pearson, 2020)

Several transaction cost mitigation strategies have been proposed: One common strategy is to optimize the holding period keeping delta-hedged options from mid-month to maturity, thereby avoiding the cost of closing the position early (Käfer, Moerke, et al., 2025; O'Donovan & Yu, 2024; Goyal & Saretto, 2022; Zhan et al., 2022). Additionally, O'Donovan and Yu (2024) suggest including only options whose quoted bid-ask spreads fall within the lowest four deciles before constructing factor portfolios. They also propose eliminating the costly short leg and instead neutralizing market volatility exposure by shorting a small position in an index option instead. Alternatively, Muravyev and Pearson (2020) argue that intraday trading of options can exploit periods when spreads are narrower than typically assumed, thereby reducing transaction costs.

Linear and nonlinear ensemble strategies currently produce less negative returns. Reducing transaction costs through the discussed strategies could push these returns of these methods into positive territory. Previous research demonstrates that delta-hedged option investments can generate positive returns after accounting for transaction costs (see, e.g., Bali et al. (2023), Zhan et al. (2022)); hence, it is a promising prospect that factor timing can also achieve positive risk-adjusted returns net

of transaction costs.

Training periods. I also test the use of expanding versus rolling window forecasts. Compared to the default rolling window forecast using 60 months of data, an expanding window forecast significantly reduces the prediction quality of the models. Table 13 shows that no advanced model achieves better performance on any PC than the benchmark of zero returns. The only model that partially benefits from using an expanding window is the ARMA model, which achieves a higher average return than before.

Model	PC1	PC2	PC3	PC4	PC5	PC6
arima	-0.02	0.41	0.02	0.09	0.04	0.03
zero_return	-0.00	0.31	0.09	0.04	0.11	0.01
lasso	-0.18	-1.33	-5.42	-0.59	-3.25	-1.05
ridge	-0.22	-3.39	-3.95	-0.60	-3.03	-0.89
en	-0.16	-1.63	-5.28	-0.56	-3.08	-1.04
pls	-0.16	-2.44	-6.01	-1.29	-3.67	-2.77
linear_ens	-0.10	-2.12	-5.10	-0.63	-3.17	-1.32
rand_forest	-0.12	-0.45	-0.66	-0.09	-0.54	-0.40
XGBoost	-0.28	-3.67	-3.68	-0.65	-2.57	-1.27
Dart	-0.26	-3.06	-3.70	-0.46	-2.67	-1.14
nonlinear_ens	-0.18	-2.13	-2.38	-0.22	-1.73	-0.79

Table 13: Out-of-sample R^2 for each model using expanding windows.

One reason for the poor performance could be structural breaks in the data.[16] Structural breaks are changes in the underlying data patterns that affect model predictions. Therefore, training the models on outdated relationships between characteristics and returns may distort predictions (Hansen, 2001). One observation supports this: squared errors for all PCs increase dramatically after 2015, hinting at a potential structural break during the GFC. This is further suggested by the sharp increase in returns after 2015 when using a rolling window. More details on the returns and errors can be found in Appendix E.2.

To isolate the influence of the GFC, I rerun the predictions, excluding the years 2008 and 2009. As shown in Table E1, the forecasting accuracy remains similar to my base case. Figure E4 shows that the GFC itself does not have a significant impact on the final returns when using a rolling window.

[16] I find evidence for structural breaks in PC returns. Table 3 shows that we reject the null hypothesis of no structural breaks of the (Utts, 1982) Rainbow test for most PCs at 95% confidence level.

Option types. Bondarenko (2014) finds that historical out-of-the-money S&P 500 put options carry a premium. To check whether there is a difference in factor timing performance between put and call options, I run the analysis for both option types separately. The R^2_{OOS} for factor return prediction and return profile are presented in Appendix E.3.

Generally, the prediction patterns for calls and puts are similar. For both option types, the prediction is positive for the first and third PCs. I achieve positive Sharpe Ratios for both option types. However, the R^2_{OOS} values are higher for call options, indicating that their PCs are easier to predict. Nevertheless, since the factor returns of put options are generally higher (i.e. the equal-weighted portfolio return is larger), call factor timing portfolios have lower Sharpe Ratios than put option portfolios.

Selection of factors. Selecting the factors with the highest in-sample PC regression fit increases the Sharpe Ratios of the factor timing portfolios. I select the five factors with the highest adjusted in-sample R^2_{OOS} in the PC regression. Then I only use their predicted factor returns for the Sharpe Ratio optimization. I also use the Ledoit and Wolf (2022) linear shrinkage of the covariance matrix, as the number of dimensions is reduced and noise likely is lower. Results are shown in Table E6. While the Sharpe Ratio of the equal-weighted portfolio goes down, it goes up for factor timing portfolios. Hence, selecting the factors with the highest fit can improve forecasting performance and economic results.

Characteristics construction. To isolate the effect of Kagkadis et al. (2024) approach, I leave out the dimensionality reduction of the characteristics. Table E7 shows the prediction results for factor return PCs. The R^2_{OOS} values do not change much overall. Nonlinear models like Dart and XGBoost perform better. Linear models, especially Lasso and Elastic Net, perform worse. Table E9 presents the final result of portfolio construction. The final Sharpe Ratios increase for linear models and decrease for nonlinear ones. Overall, I do not observe a significant marginal benefit from performing dimensionality reduction with the characteristics, raising the question of effectiveness for the method of applying PCA to the characteristics for options factors.

I also check whether the characteristics data improves prediction quality generally. Therefore, I exclude the characteristics, leaving only the non-tradable option factors, the overall predictors, and the momentum factors of PCs as explanatory

variables. Table E8 shows the results. While some linear models like Lasso and PLS perform better, others like Ridge and EN show lower accuracy. Nonlinear models improve slightly. Table E10 presents the final economic results. Excluding the characteristics matrix from the explanatory variables reduces performance for almost every model, except XGBoost. This suggests that including the characteristics is generally beneficial, corresponding with the finding of Kagkadis et al. (2024). Similar to equity factors, the characteristics of option factors themselves contain predictive power for factor returns. This finding suggests that anomalies are not purely statistical, but are linked to economically interpretable firm and option market attributes.

Data snooping. Both out-of-sample forecast evaluation and cross-validation are methods that I use to reduce the risk of overfitting. Yet, evaluating many forecasts increases the likelihood of data snooping (Timmermann, 2018; Giglio, Liao, & Xiu, 2021). Data snooping is the practice of formulating a hypothesis after looking at the data, which results in invalid statistical inference. White (2000) proposes a method to address this issue. I run the White (2000) test using stationary bootstrap to preserve the time series structure of the data. The key difference from the Clark and West (2007) approach defined in Equations 7 and 8 is that I use bootstrapping for each PC and across models, whereas Clark and West (2007) evaluates the forecasts across PCs for each model.

Model	**1**	**2**	**3**	**4**	**5**	**6**
lasso	2.06	0.08	-0.12	0.00	-0.07	-0.20
ridge	1.15	0.43	0.08	0.04	0.01	-0.01
en	2.23	0.08	-0.03	-0.05	-0.10	-0.18
pls	0.61	-0.11	-0.29	-0.21	-0.31	-0.24
linear_ens	2.17	0.32	0.02	0.05	0.02	-0.10
arima	0.18	0.37	-0.10	-0.00	-0.01	0.03
nonlinear_ens	1.80	0.34	-0.03	-0.04	-0.06	-0.08
XGBoost	1.72	0.21	-0.14	-0.06	-0.13	-0.15
Dart	1.84	0.35	-0.04	-0.08	-0.08	-0.07
rand_forest	1.23	0.23	-0.07	-0.13	-0.11	-0.14
p-value	0.00	0.16	0.50	0.61	0.82	0.76

Table 14: This table shows then average loss differential $\bar{d}^m$ and the p-value of the White (2000) reality check test. A positive loss differential $\bar{d}^m > 0$ means that the model is on average outperforming the benchmark (historical average). I use the squared loss function, 1000 bootstraps, and a block length of 10. The p-value is defined as the proportion of bootstrapped maxima that exceed or equal the observed.

Table 14 presents the results, confirming key observations made earlier. ML models achieve statistically significant forecasting performance for the first PC. Still,

forecasting subsequent PCs remains a challenge, with all other PCs not being statistically significantly estimated, with a p-value of more than 16%. This is despite the positive loss differential indicating superior forecasting performance of almost all models.

Look-ahead bias can lead to high mean returns and Sharpe Ratios for option trading strategies. Duarte, Jones, Khorram, and Mo (2023) point out that using *ex post* information to filter option return data can lead to this bias. As the option filter is applied only at position initiation, following Bali et al. (2023), the issue of filtering does not apply to my data. However, I use predictors that are not available during the entire sample period. For example, the sentiment index of Baker and Wurgler (2006) was not available before the publication of the paper, possibly inducing look-ahead bias.

6 Conclusion

At the highest level, my research finds that option factor returns are, in fact, predictable and that ML models can generate economically meaningful value.

My work has theoretical implications for future research. On the one hand, I find that the cross-sectional variation in factor returns can largely be explained by a few linear components. In this sense, option factors behave similarly to equity factors, as both display a low-dimensional linear structure. On the other hand, when focusing on the time-series dimension, I find that the factor exposures in an optimal factor timing portfolio vary considerably. Because factor risk compensation changes over time, these shifting weights aim to capture predictable patterns in premia. Here, no single characteristic dominates, and using a broad set of predictors is beneficial. Hence, forecasting time-series return patterns is a high-dimensional and dynamic problem. Taken together, these findings imply that while the factor structure is linear and low-dimensional in the cross-section, the dynamics of risk compensation are complex. These observations suggest that the SDF is dynamic and state-dependent: it incorporates evolving market conditions, volatility, and investor preferences.

Yet, three key limitations to my work should be mentioned.

First, option factor timing portfolios do not outperform an equally weighted static factor portfolio in terms of risk-adjusted returns, due to issues associated with covariance estimation and prediction. Further research should focus on implement-

ing similar methods to Kozak, Nagel, and Santosh (2020) to optimize the covariance of large-dimensional return series for option factor returns. For example, one could apply non-linear shrinkage to increase accuracy or incorporate dynamic development of covariance instead of using a static estimate.

Second, another problem with these option factor timing portfolios is transaction costs. Applying a proxy of about 20% of the half bid-ask spread results in negative returns. My research is limited by the data set, that only contains data from beginning of the month to month end. Future research must account for transaction cost optimization. For instance, holding options until maturity or filtering for low-cost options could increase net returns further (O'Donovan & Yu, 2024). As previous research proves that delta-hedged option investing can be profitable, the application of transaction cost mitigation strategies is promising.

Third, my forecasting methods lack interpretability. Features vary significantly across models and PCs. This makes identifying performance drivers difficult. Section 5.3 explains this issue. Other papers could discuss feature importance more thoroughly, covering permutation feature importance, time-varying feature analysis, and the role of macroeconomic predictors, similar to Gu et al. (2020). Alternatively, SHAP values could help interpret model decisions (Lundberg & Lee, 2017).

Overall, my results indicate a trade-off between forecasting accuracy and portfolio implementation in option factor markets. While novel ML methods can predict return patterns, turning these forecasts into profits is difficult, particularly when considering transaction costs and risk adjustments.

References

Agarwal, V., Arisoy, Y. E., & Naik, N. Y. (2017, September). Volatility of aggregate volatility and hedge fund returns. *Journal of Financial Economics*, *125*(3), 491–510. Retrieved 2025-08-10, from https://linkinghub.elsevier.com/retrieve/pii/S0304405X17301320 doi: 10.1016/j.jfineco.2017.06.015

An, B., Ang, A., Bali, T. G., & Cakici, N. (2014, October). The Joint Cross Section of Stocks and Options. *The Journal of Finance*, *69*(5), 2279–2337. Retrieved 2025-05-25, from https://onlinelibrary.wiley.com/doi/10.1111/jofi.12181 doi: 10.1111/jofi.12181

Aretz, K., Lin, M.-T., & Poon, S.-H. (2023, January). Moneyness, Underlying Asset Volatility, and the Cross-Section of Option Returns. *Review of Finance*, *27*(1), 289–323. Retrieved 2025-06-12, from https://academic.oup.com/rof/article/27/1/289/6510952 doi: 10.1093/rof/rfac003

Asness, C., Chandra, S., Ilmanen, A., & Israel, R. (2017, March). Contrarian Factor Timing is Deceptively Difficult. *The Journal of Portfolio Management*, *43*(5), 72–87. Retrieved 2025-05-25, from http://pm-research.com/lookup/doi/10.3905/jpm.2017.43.5.072 doi: 10.3905/jpm.2017.43.5.072

Baker, M., & Wurgler, J. (2006, August). Investor Sentiment and the Cross-Section of Stock Returns. *The Journal of Finance*, *61*(4), 1645–1680. Retrieved 2025-07-01, from https://onlinelibrary.wiley.com/doi/10.1111/j.1540-6261.2006.00885.x doi: 10.1111/j.1540-6261.2006.00885.x

Bakshi, G., & Kapadia, N. (2003, April). Delta-Hedged Gains and the Negative Market Volatility Risk Premium. *Review of Financial Studies*, *16*(2), 527–566. Retrieved 2025-04-10, from https://academic.oup.com/rfs/article-lookup/doi/10.1093/rfs/hhg002 doi: 10.1093/rfs/hhg002

Bali, T. G., Beckmeyer, H., Mörke, M., & Weigert, F. (2023, August). Option Return Predictability with Machine Learning and Big Data. *The Review of Financial Studies*, *36*(9), 3548–3602. Retrieved 2025-04-11, from https://academic.oup.com/rfs/article/36/9/3548/7056660 doi: 10.1093/rfs/hhad017

Bali, T. G., & Murray, S. (2013, August). Does Risk-Neutral Skewness Predict the Cross-Section of Equity Option Portfolio Returns? *Journal of Financial and Quantitative Analysis*, *48*(4), 1145–1171. Retrieved 2025-05-25, from https://www.cambridge.org/core/product/identifier/S0022109013000410/type/journal_article doi: 10.1017/S0022109013000410

Bates, J. M., & Granger, C. W. J. (1969, December). The Combination of Forecasts. *OR*, *20*(4), 451. Retrieved 2025-05-07, from https://www.jstor.org/stable/3008764?origin=crossref doi: 10.2307/3008764

Bauwens, L., Laurent, S., & Rombouts, J. V. K. (2006, January). Multivariate GARCH models: a survey. *Journal of Applied Econometrics*, *21*(1), 79–109. Re-

trieved 2025-06-11, from https://onlinelibrary.wiley.com/doi/10.1002/jae.842 doi: 10.1002/jae.842

Black, F., & Scholes, M. (1973, May). The Pricing of Options and Corporate Liabilities. *Journal of Political Economy*, *81*(3), 637–654. Retrieved 2025-05-25, from https://www.journals.uchicago.edu/doi/10.1086/260062 doi: 10.1086/260062

Bondarenko, O. (2014, September). Why Are Put Options So Expensive? *Quarterly Journal of Finance*, *04*(03), 1450015. Retrieved 2025-05-25, from https://www.worldscientific.com/doi/abs/10.1142/S2010139214500153 doi: 10.1142/S2010139214500153

Boulatov, A., Eisdorfer, A., Goyal, A., & Zhdanov, A. (2022). Limited Attention and Option Prices. *SSRN Electronic Journal*. Retrieved 2025-06-12, from https://www.ssrn.com/abstract=3607030 doi: 10.2139/ssrn.3607030

Boyer, B. H., & Vorkink, K. (2014, August). Stock Options as Lotteries. *The Journal of Finance*, *69*(4), 1485–1527. Retrieved 2025-05-25, from https://onlinelibrary.wiley.com/doi/10.1111/jofi.12152 doi: 10.1111/jofi.12152

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. Retrieved 2025-05-20, from http://link.springer.com/10.1023/A:1010933404324 doi: 10.1023/A:1010933404324

Bryzgalova, S., Huang, J., & Julliard, C. (2023, February). Bayesian Solutions for the Factor Zoo: We Just Ran Two Quadrillion Models. *The Journal of Finance*, *78*(1), 487–557. Retrieved 2025-05-24, from https://onlinelibrary.wiley.com/doi/10.1111/jofi.13197 doi: 10.1111/jofi.13197

Buja, A., & Eyuboglu, N. (1992, October). Remarks on Parallel Analysis. *Multivariate Behavioral Research*, *27*(4), 509–540. Retrieved 2025-04-10, from http://www.tandfonline.com/doi/abs/10.1207/s15327906mbr2704_2 doi: 10.1207/s15327906mbr2704_2

Buraschi, A., Kosowski, R., & Trojani, F. (2014, February). When There Is No Place to Hide: Correlation Risk and the Cross-Section of Hedge Fund Returns. *Review of Financial Studies*, *27*(2), 581–616. Retrieved 2025-08-10, from https://academic.oup.com/rfs/article-lookup/doi/10.1093/rfs/hht070 doi: 10.1093/rfs/hht070

Byun, S.-J., & Kim, D.-H. (2016, October). Gambling preference and individual equity option returns. *Journal of Financial Economics*, *122*(1), 155–174. Retrieved 2025-05-25, from https://linkinghub.elsevier.com/retrieve/pii/S0304405X1630109X doi: 10.1016/j.jfineco.2016.06.004

Büchner, M., & Kelly, B. (2022, March). A factor model for option returns. *Journal of Financial Economics*, *143*(3), 1140–1161. Retrieved 2025-05-24, from https://linkinghub.elsevier.com/retrieve/pii/S0304405X21005249 doi: 10

.1016/j.jfineco.2021.12.007

Cao, J., & Han, B. (2013, April). Cross section of option returns and idiosyncratic stock volatility. *Journal of Financial Economics*, *108*(1), 231–249. Retrieved 2025-05-25, from https://linkinghub.elsevier.com/retrieve/pii/S0304405X12002450 doi: 10.1016/j.jfineco.2012.11.010

Cenesizoglu, T., & Timmermann, A. (2012, November). Do return prediction models add economic value? *Journal of Banking & Finance*, *36*(11), 2974–2987. Retrieved 2025-07-27, from https://linkinghub.elsevier.com/retrieve/pii/S0378426612001604 doi: 10.1016/j.jbankfin.2012.06.008

Chen, L., Pelger, M., & Zhu, J. (2024, February). Deep Learning in Asset Pricing. *Management Science*, *70*(2), 714–750. Retrieved 2025-05-25, from https://pubsonline.informs.org/doi/10.1287/mnsc.2023.4695 doi: 10.1287/mnsc.2023.4695

Chen, T., & Guestrin, C. (2016, August). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). San Francisco California USA: ACM. Retrieved 2025-06-09, from https://dl.acm.org/doi/10.1145/2939672.2939785 doi: 10.1145/2939672.2939785

Chong, I.-G., & Jun, C.-H. (2005, July). Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory Systems*, *78*(1-2), 103–112. Retrieved 2025-06-09, from https://linkinghub.elsevier.com/retrieve/pii/S0169743905000031 doi: 10.1016/j.chemolab.2004.12.011

Christoffersen, P., Fournier, M., & Jacobs, K. (2018, February). The Factor Structure in Equity Options. *The Review of Financial Studies*, *31*(2), 595–637. Retrieved 2025-05-25, from https://academic.oup.com/rfs/article/31/2/595/4060546 doi: 10.1093/rfs/hhx089

Christoffersen, P., Goyenko, R., Jacobs, K., & Karoui, M. (2018, March). Illiquidity Premia in the Equity Options Market. *The Review of Financial Studies*, *31*(3), 811–851. Retrieved 2025-05-25, from https://academic.oup.com/rfs/article/31/3/811/4371415 doi: 10.1093/rfs/hhx113

Clark, T. E., & West, K. D. (2007, May). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, *138*(1), 291–311. Retrieved 2025-04-11, from https://linkinghub.elsevier.com/retrieve/pii/S0304407606000960 doi: 10.1016/j.jeconom.2006.05.023

Cleveland, W. S., & Devlin, S. J. (1988, September). Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association*, *83*(403), 596–610. Retrieved 2025-05-21, from http://www.tandfonline.com/doi/abs/10.1080/01621459.1988.10478639 doi: 10.1080/01621459.1988.10478639

Cochrane, J. H. (2011, August). Presidential Address: Discount Rates. *The Journal of Finance*, *66*(4), 1047–1108. Retrieved 2025-07-20, from https://onlinelibrary.wiley.com/doi/10.1111/j.1540-6261.2011.01671.x (Publisher: Wiley) doi: 10.1111/j.1540-6261.2011.01671.x

Constantinides, G. M., Jackwerth, J. C., & Savov, A. (2013, December). The Puzzle of Index Option Returns. *Review of Asset Pricing Studies*, *3*(2), 229–257. Retrieved 2025-05-25, from https://academic.oup.com/raps/article-lookup/doi/10.1093/rapstu/rat004 doi: 10.1093/rapstu/rat004

Coval, J. D., & Shumway, T. (2001, June). Expected Option Returns. *The Journal of Finance*, *56*(3), 983–1009. Retrieved 2025-05-25, from https://onlinelibrary.wiley.com/doi/10.1111/0022-1082.00352 doi: 10.1111/0022-1082.00352

Dayal, B. S., & MacGregor, J. F. (1997, January). Improved PLS algorithms. *Journal of Chemometrics*, *11*(1), 73–85. Retrieved from http://dx.doi.org/10.1002/(sici)1099-128x(199701)11:1<73::aid-cem435>3.0.co;2-# (Publisher: Wiley) doi: 10.1002/(sici)1099-128x(199701)11:1<73::aid-cem435>3.0.co;2-#

Detzel, A., Novy-Marx, R., & Velikov, M. (2023, June). Model Comparison with Transaction Costs. *The Journal of Finance*, *78*(3), 1743–1775. Retrieved 2025-07-23, from https://onlinelibrary.wiley.com/doi/10.1111/jofi.13225 (Publisher: Wiley) doi: 10.1111/jofi.13225

Didisheim, A., Ke, S. B., Kelly, B., & Malamud, S. (2024, September). *APT or "AIPT"? The Surprising Dominance of Large Factor Models* (Tech. Rep. No. w33012). Cambridge, MA: National Bureau of Economic Research. Retrieved 2025-06-06, from http://www.nber.org/papers/w33012.pdf doi: 10.3386/w33012

Diebold, F. X., & Mariano, R. S. (1995, July). Comparing Predictive Accuracy. *Journal of Business & Economic Statistics*, *13*(3), 253–263. Retrieved 2025-04-11, from http://www.tandfonline.com/doi/abs/10.1080/07350015.1995.10524599 doi: 10.1080/07350015.1995.10524599

Duarte, J., Jones, C. S., Khorram, M., & Mo, H. (2023). Too Good to Be True: Look-ahead Bias in Empirical Option Research. *SSRN Electronic Journal*. Retrieved 2025-06-19, from https://www.ssrn.com/abstract=4590083 doi: 10.2139/ssrn.4590083

Duarte, J., Jones, C. S., & Wang, J. L. (2024, October). Very Noisy Option Prices and Inference Regarding the Volatility Risk Premium. *The Journal of Finance*, *79*(5), 3581–3621. Retrieved 2025-05-25, from https://onlinelibrary.wiley.com/doi/10.1111/jofi.13365 doi: 10.1111/jofi.13365

Ehsani, S., & Linnainmaa, J. T. (2022, June). Factor Momentum and the Momentum Factor. *The Journal of Finance*, *77*(3), 1877–1919. Retrieved 2025-05-24, from https://onlinelibrary.wiley.com/doi/10.1111/jofi.13131 doi: 10.1111/jofi.13131

Fabrigar, L. R., & Wegener, D. T. (2011). *Exploratory Factor Analysis*. Oxford University Press. Retrieved 2025-08-24, from https://academic.oup.com/book/41736 doi: 10.1093/acprof:osobl/9780199734177.001.0001

Fama, E. F., & French, K. R. (1993, February). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, *33*(1), 3–56. Retrieved 2025-06-06, from https://linkinghub.elsevier.com/retrieve/pii/0304405X93900235 doi: 10.1016/0304-405X(93)90023-5

Fama, E. F., & French, K. R. (2015, April). A five-factor asset pricing model. *Journal of Financial Economics*, *116*(1), 1–22. Retrieved 2025-07-01, from https://linkinghub.elsevier.com/retrieve/pii/S0304405X14002323 doi: 10.1016/j.jfineco.2014.10.010

Feng, G., Giglio, S., & Xiu, D. (2020, June). Taming the Factor Zoo: A Test of New Factors. *The Journal of Finance*, *75*(3), 1327–1370. Retrieved 2025-05-24, from https://onlinelibrary.wiley.com/doi/10.1111/jofi.12883 doi: 10.1111/jofi.12883

Frazzini, A., & Pedersen, L. H. (2022, February). Embedded Leverage. *The Review of Asset Pricing Studies*, *12*(1), 1–52. Retrieved 2025-06-06, from https://academic.oup.com/raps/article/12/1/1/6373911 doi: 10.1093/rapstu/raab022

Freyberger, J., Neuhierl, A., & Weber, M. (2020, May). Dissecting Characteristics Nonparametrically. *The Review of Financial Studies*, *33*(5), 2326–2377. Retrieved 2025-05-25, from https://academic.oup.com/rfs/article/33/5/2326/5821383 doi: 10.1093/rfs/hhz123

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, *33*(1). Retrieved 2025-05-07, from http://www.jstatsoft.org/v33/i01/ doi: 10.18637/jss.v033.i01

Friedman, J. H. (2001, October). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5). Retrieved 2025-05-20, from https://projecteuclid.org/journals/annals-of-statistics/volume-29/issue-5/Greedy-function-approximation-A-gradient-boosting-machine/10.1214/aos/1013203451.full doi: 10.1214/aos/1013203451

Giglio, S., Liao, Y., & Xiu, D. (2021, June). Thousands of Alpha Tests. *The Review of Financial Studies*, *34*(7), 3456–3496. Retrieved 2025-05-25, from https://academic.oup.com/rfs/article/34/7/3456/5911131 doi: 10.1093/rfs/hhaa111

Gourio, F. (2013, July). Credit Risk and Disaster Risk. *American Economic Journal: Macroeconomics*, *5*(3), 1–34. Retrieved 2025-07-20, from https://pubs.aeaweb.org/doi/10.1257/mac.5.3.1 (Publisher: American Economic Association) doi: 10.1257/mac.5.3.1

Goyal, A., & Saretto, A. (2009, November). Cross-section of option returns and volatility. *Journal of Financial Economics*, *94*(2), 310–326. Retrieved 2025-05-25, from https://linkinghub.elsevier.com/retrieve/pii/S0304405X09001251 doi: 10.1016/j.jfineco.2009.01.001

Goyal, A., & Saretto, A. (2022, August). Are Equity Option Returns Abnormal? IPCA Says No. *Federal Reserve Bank of Dallas, Working Papers*, *2022*(2214). Retrieved 2025-07-23, from https://www.dallasfed.org/-/media/documents/research/papers/2022/wp2214.pdf (Publisher: Federal Reserve Bank of Dallas) doi: 10.24149/wp2214

Goyal, A., & Saretto, A. (2025, June). Can Equity Option Returns Be Explained by a Factor Model? IPCA Says Yes. *The Review of Financial Studies*, *38*(6), 1783–1821. Retrieved 2025-05-25, from https://academic.oup.com/rfs/article/38/6/1783/8010873 doi: 10.1093/rfs/hhae087

Goyal, A., Welch, I., & Zafirov, A. (2024, November). A Comprehensive 2022 Look at the Empirical Performance of Equity Premium Prediction. *The Review of Financial Studies*, *37*(11), 3490–3557. Retrieved 2025-04-10, from https://academic.oup.com/rfs/article/37/11/3490/7749383 doi: 10.1093/rfs/hhae044

Goyenko, R., & Zhang, C. (2020). The Joint Cross Section of Option and Stock Returns Predictability with Big Data and Machine Learning. *SSRN Electronic Journal*. Retrieved 2025-05-25, from https://www.ssrn.com/abstract=3747238 doi: 10.2139/ssrn.3747238

Grammig, J., Hanenberg, C., Schlag, C., & Sönksen, J. (2025, January). Diverging Roads: Theory-Based vs. Machine Learning-Implied Stock Risk Premia. *Journal of Financial Econometrics*, *23*(2), nbaf005. Retrieved 2025-05-25, from https://academic.oup.com/jfec/article/doi/10.1093/jjfinec/nbaf005/8074497 doi: 10.1093/jjfinec/nbaf005

Gu, S., Kelly, B., & Xiu, D. (2020, May). Empirical Asset Pricing via Machine Learning. *The Review of Financial Studies*, *33*(5), 2223–2273. Retrieved 2025-05-24, from https://academic.oup.com/rfs/article/33/5/2223/5758276 doi: 10.1093/rfs/hhaa009

Gu, S., Kelly, B., & Xiu, D. (2021, May). Autoencoder asset pricing models. *Journal of Econometrics*, *222*(1), 429–450. Retrieved 2025-05-24, from https://linkinghub.elsevier.com/retrieve/pii/S0304407620301998 doi: 10.1016/j.jeconom.2020.07.009

Gupta, T., & Kelly, B. (2019, February). Factor Momentum Everywhere. *The Journal of Portfolio Management*, *45*(3), 13–36. Retrieved 2025-05-24, from http://pm-research.com/lookup/doi/10.3905/jpm.2019.45.3.013 doi: 10.3905/jpm.2019.45.3.013

Haddad, V., Kozak, S., & Santosh, S. (2020, May). Factor Timing. *The Review*

of Financial Studies, *33*(5), 1980–2018. Retrieved 2025-04-10, from https://academic.oup.com/rfs/article/33/5/1980/5753962 doi: 10.1093/rfs/hhaa017

Hansen, B. E. (2001, November). The New Econometrics of Structural Change: Dating Breaks in U.S. Labor Productivity. *Journal of Economic Perspectives*, *15*(4), 117–128. Retrieved 2025-06-07, from https://pubs.aeaweb.org/doi/10.1257/jep.15.4.117 doi: 10.1257/jep.15.4.117

Harvey, A. C., & Collier, P. (1977, July). Testing for functional misspecification in regression analysis. *Journal of Econometrics*, *6*(1), 103–119. Retrieved 2025-05-20, from https://linkinghub.elsevier.com/retrieve/pii/0304407677900574 doi: 10.1016/0304-4076(77)90057-4

He, Z., Kelly, B., & Manela, A. (2017, October). Intermediary asset pricing: New evidence from many asset classes. *Journal of Financial Economics*, *126*(1), 1–35. Retrieved 2025-08-10, from https://linkinghub.elsevier.com/retrieve/pii/S0304405X1730212X doi: 10.1016/j.jfineco.2017.08.002

Heston, S. L., Jones, C. S., Khorram, M., Li, S., & Mo, H. (2023, December). Option Momentum. *The Journal of Finance*, *78*(6), 3141–3192. Retrieved 2025-05-24, from https://onlinelibrary.wiley.com/doi/10.1111/jofi.13279 doi: 10.1111/jofi.13279

Ho, T., Kagkadis, A., & Wang, G. (2024, February). Is Firm-Level Political Risk Priced in the Equity Option Market? *The Review of Asset Pricing Studies*, *14*(1), 153–195. Retrieved 2025-05-25, from https://academic.oup.com/raps/article/14/1/153/7331038 doi: 10.1093/rapstu/raad013

Hoerl, A. E., & Kennard, R. W. (1970, February). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, *12*(1), 55–67. Retrieved 2025-05-07, from http://www.tandfonline.com/doi/abs/10.1080/00401706.1970.10488634 doi: 10.1080/00401706.1970.10488634

Horenstein, A. R., Vasquez, A., & Xiao, X. (2023). Common Factors in Equity Option Returns. *SSRN Electronic Journal*. Retrieved 2025-05-24, from https://www.ssrn.com/abstract=3290363 doi: 10.2139/ssrn.3290363

Horn, J. L. (1965, June). A Rationale and Test for the Number of Factors in Factor Analysis. *Psychometrika*, *30*(2), 179–185. Retrieved 2025-04-10, from https://www.cambridge.org/core/product/identifier/S003331230004165X/type/journal_article doi: 10.1007/BF02289447

Hu, G., & Jacobs, K. (2020, May). Volatility and Expected Option Returns. *Journal of Financial and Quantitative Analysis*, *55*(3), 1025–1060. Retrieved 2025-05-25, from https://www.cambridge.org/core/product/identifier/S0022109019000310/type/journal_article doi: 10.1017/S0022109019000310

Hyndman, R. J., & Khandakar, Y. (2008). Automatic Time Series Forecasting: The

forecast Package for *R*. *Journal of Statistical Software*, *27*(3). Retrieved 2025-05-08, from http://www.jstatsoft.org/v27/i03/ doi: 10.18637/jss.v027.i03

Jarque, C. M., & Bera, A. K. (1980, January). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, *6*(3), 255–259. Retrieved 2025-05-21, from https://linkinghub.elsevier.com/retrieve/pii/0165176580900245 doi: 10.1016/0165-1765(80)90024-5

Jeon, Y., Kan, R., & Li, G. (2025, June). Stock Return Autocorrelations and Expected Option Returns. *Management Science*, *71*(6), 4895–4914. Retrieved 2025-06-12, from https://pubsonline.informs.org/doi/10.1287/mnsc.2023.03071 doi: 10.1287/mnsc.2023.03071

Jolliffe, I. (2002). *Principal Component Analysis*. New York: Springer-Verlag. Retrieved 2025-08-31, from http://link.springer.com/10.1007/b98835 doi: 10.1007/b98835

Jurado, K., Ludvigson, S. C., & Ng, S. (2015, March). Measuring Uncertainty. *American Economic Review*, *105*(3), 1177–1216. Retrieved 2025-08-10, from https://pubs.aeaweb.org/doi/10.1257/aer.20131193 doi: 10.1257/aer.20131193

Kagkadis, A., Nolte, I., Nolte, S., & Vasilas, N. (2024, February). Factor Timing with Portfolio Characteristics. *The Review of Asset Pricing Studies*, *14*(1), 84–118. Retrieved 2025-04-10, from https://academic.oup.com/raps/article/14/1/84/7191017 doi: 10.1093/rapstu/raad010

Kanne, S., Korn, O., & Uhrig-Homburg, M. (2023, March). Stock illiquidity and option returns. *Journal of Financial Markets*, *63*, 100765. Retrieved 2025-05-25, from https://linkinghub.elsevier.com/retrieve/pii/S1386418122000556 doi: 10.1016/j.finmar.2022.100765

Kelly, B., & Xiu, D. (2023, July). *Financial Machine Learning* (Tech. Rep. No. w31502). Cambridge, MA: National Bureau of Economic Research. Retrieved 2025-05-08, from http://www.nber.org/papers/w31502.pdf doi: 10.3386/w31502

Kelly, B. T., Pruitt, S., & Su, Y. (2019, December). Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics*, *134*(3), 501–524. Retrieved 2025-06-13, from https://linkinghub.elsevier.com/retrieve/pii/S0304405X19301151 doi: 10.1016/j.jfineco.2019.05.001

Kozak, S., Nagel, S., & Santosh, S. (2020, February). Shrinking the cross-section. *Journal of Financial Economics*, *135*(2), 271–292. Retrieved 2025-05-25, from https://linkinghub.elsevier.com/retrieve/pii/S0304405X19301655 doi: 10.1016/j.jfineco.2019.06.008

Käfer, N. (2025). Options on Drugs: Industry Exposure and Option Anomalies. *SSRN Electronic Journal*. Retrieved 2025-06-13, from https://www.ssrn.com/abstract=5245162 doi: 10.2139/ssrn.5245162

Käfer, N., Moerke, M., Weigert, F., & Wiest, T. (2025). A Bayesian SDF

for Equity Options. *Journal of Financial and Quantitative Analysis (forthcoming)*. Retrieved 2025-09-28, from https://jfqa.org/2025/08/01/a-bayesian-stochastic-discount-factor-for-the-cross-section-of-individual-equity-options/ doi: 10.2139/ssrn.4710335

Käfer, N., Mörke, M., & Wiest, T. (2025, April). Option Factor Momentum. *Journal of Financial and Quantitative Analysis*, 1–52. Retrieved 2025-05-24, from https://www.cambridge.org/core/product/identifier/S0022109025000225/type/journal_article doi: 10.1017/S0022109025000225

Ledoit, O., & Wolf, M. (2008, December). Robust performance hypothesis testing with the Sharpe ratio. *Journal of Empirical Finance*, *15*(5), 850–859. Retrieved 2025-05-20, from https://linkinghub.elsevier.com/retrieve/pii/S0927539808000182 doi: 10.1016/j.jempfin.2008.03.002

Ledoit, O., & Wolf, M. (2022, January). The Power of (Non-)Linear Shrinking: A Review and Guide to Covariance Matrix Estimation. *Journal of Financial Econometrics*, *20*(1), 187–218. Retrieved 2025-06-11, from https://academic.oup.com/jfec/article/20/1/187/5861007 doi: 10.1093/jjfinec/nbaa007

Lehnherr, R., Mehta, M., & Nagel, S. (2025, April). Optimal Factor Timing in a High-Dimensional Setting. *Financial Analysts Journal*, *81*(2), 51–66. Retrieved 2025-05-25, from https://www.tandfonline.com/doi/full/10.1080/0015198X.2025.2474385 doi: 10.1080/0015198X.2025.2474385

Lettau, M., & Ludvigson, S. (2001, June). Consumption, Aggregate Wealth, and Expected Stock Returns. *The Journal of Finance*, *56*(3), 815–849. Retrieved 2025-07-01, from https://onlinelibrary.wiley.com/doi/10.1111/0022-1082.00347 doi: 10.1111/0022-1082.00347

Lundberg, S., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions.* arXiv. Retrieved 2025-07-27, from https://arxiv.org/abs/1705.07874 (Version Number: 2) doi: 10.48550/ARXIV.1705.07874

Ma, T., Liao, C., & Jiang, F. (2023, March). Timing the factor zoo via deep learning: Evidence from China. *Accounting & Finance*, *63*(1), 485–505. Retrieved 2025-06-12, from https://onlinelibrary.wiley.com/doi/10.1111/acfi.13033 doi: 10.1111/acfi.13033

Martin, I. W., & Nagel, S. (2022, July). Market efficiency in the age of big data. *Journal of Financial Economics*, *145*(1), 154–177. Retrieved 2025-05-25, from https://linkinghub.elsevier.com/retrieve/pii/S0304405X21004566 doi: 10.1016/j.jfineco.2021.10.006

Messmer, M., & Audrino, F. (2022, November). The Lasso and the Factor Zoo-Predicting Expected Returns in the Cross-Section. *Forecasting*, *4*(4), 969–1003. Retrieved 2025-05-24, from https://www.mdpi.com/2571-9394/4/4/53 doi: 10.3390/forecast4040053

Mevik, B.-H., & Wehrens, R. (2007). The pls Package: Principal Component and Partial Least Squares Regression in R. *Journal of Statistical Software*, *18*(2). Retrieved 2025-05-07, from http://www.jstatsoft.org/v18/i02/ doi: 10.18637/jss.v018.i02

Moreira, A., & Muir, T. (2017, August). Volatility-Managed Portfolios. *The Journal of Finance*, *72*(4), 1611–1644. Retrieved 2025-05-25, from https://onlinelibrary.wiley.com/doi/10.1111/jofi.12513 doi: 10.1111/jofi.12513

Muravyev, D. (2016, April). Order Flow and Expected Option Returns. *The Journal of Finance*, *71*(2), 673–708. Retrieved 2025-05-25, from https://onlinelibrary.wiley.com/doi/10.1111/jofi.12380 doi: 10.1111/jofi.12380

Muravyev, D., & Pearson, N. D. (2020, November). Options Trading Costs Are Lower than You Think. *The Review of Financial Studies*, *33*(11), 4973–5014. Retrieved 2025-05-25, from https://academic.oup.com/rfs/article/33/11/4973/5732665 doi: 10.1093/rfs/hhaa010

Murray, S., Xia, Y., & Xiao, H. (2024, March). Charting by machines. *Journal of Financial Economics*, *153*, 103791. Retrieved 2025-05-25, from https://linkinghub.elsevier.com/retrieve/pii/S0304405X2400014X doi: 10.1016/j.jfineco.2024.103791

Neely, C. J., Rapach, D. E., Tu, J., & Zhou, G. (2014, July). Forecasting the Equity Risk Premium: The Role of Technical Indicators. *Management Science*, *60*(7), 1772–1791. Retrieved 2025-07-13, from https://pubsonline.informs.org/doi/10.1287/mnsc.2013.1838 (Publisher: Institute for Operations Research and the Management Sciences (INFORMS)) doi: 10.1287/mnsc.2013.1838

Neuhierl, A., Randl, O., Reschenhofer, C., & Zechner, J. (2023). Timing the Factor Zoo. *SSRN Electronic Journal*. Retrieved 2025-05-24, from https://www.ssrn.com/abstract=4376898 doi: 10.2139/ssrn.4376898

Newey, W. K., & West, K. D. (1987, May). A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica*, *55*(3), 703. Retrieved 2025-04-11, from https://www.jstor.org/stable/1913610?origin=crossref doi: 10.2307/1913610

O'Donovan, J., & Yu, G. Y. (2024). Transaction Costs and Cost Mitigation in Option Investment Strategies. *SSRN Electronic Journal*. Retrieved 2025-06-06, from https://www.ssrn.com/abstract=4806038 doi: 10.2139/ssrn.4806038

Ofek, E., Richardson, M., & Whitelaw, R. F. (2004, November). Limited arbitrage and short sales restrictions: evidence from the options markets. *Journal of Financial Economics*, *74*(2), 305–342. Retrieved 2025-06-06, from https://linkinghub.elsevier.com/retrieve/pii/S0304405X04000662 doi: 10.1016/j.jfineco.2003.05.008

Pafka, S., & Kondor, I. (2001, October). Evaluating the RiskMetrics methodology in measuring volatility and Value-at-Risk in financial markets. *Physica A: Statistical Mechanics and its Applications*, *299*(1-2), 305–310. Retrieved 2025-06-11, from https://linkinghub.elsevier.com/retrieve/pii/S0378437101003107 doi: 10.1016/S0378-4371(01)00310-7

Pastor, L., & Stambaught, R. F. (2003, June). Liquidity Risk and Expected Stock Returns. *Journal of Political Economy*, *111*(3), 642–685. Retrieved 2025-08-10, from https://www.journals.uchicago.edu/doi/10.1086/374184 doi: 10.1086/374184

Ramachandran, L. S., & Tayal, J. (2021, July). Mispricing, short-sale constraints, and the cross-section of option returns. *Journal of Financial Economics*, *141*(1), 297–321. Retrieved 2025-05-25, from https://linkinghub.elsevier.com/retrieve/pii/S0304405X21000969 doi: 10.1016/j.jfineco.2021.03.006

Ramsey, J. B. (1969, July). Tests for Specification Errors in Classical Linear Least-Squares Regression Analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *31*(2), 350–371. Retrieved 2025-04-10, from https://academic.oup.com/jrsssb/article/31/2/350/7027014 doi: 10.1111/j.2517-6161.1969.tb00796.x

Rapach, D. E., Ringgenberg, M. C., & Zhou, G. (2016, July). Short interest and aggregate stock returns. *Journal of Financial Economics*, *121*(1), 46–65. Retrieved 2025-07-13, from https://linkinghub.elsevier.com/retrieve/pii/S0304405X16300320 (Publisher: Elsevier BV) doi: 10.1016/j.jfineco.2016.03.004

Rashmi, K. V., & Gilad-Bachrach, R. (2015). *DART: Dropouts meet Multiple Additive Regression Trees.* arXiv. Retrieved 2025-05-20, from https://arxiv.org/abs/1505.01866 (Version Number: 1) doi: 10.48550/ARXIV.1505.01866

Risk Management in Central Banks in the Context of Monetary Policy Normalization. (2019). In *Springer Proceedings in Business and Economics* (pp. 279–289). Cham: Springer International Publishing. Retrieved 2025-07-20, from http://link.springer.com/10.1007/978-3-030-16045-6_14 (ISSN: 2198-7246, 2198-7254) doi: 10.1007/978-3-030-16045-6_14

Rosipal, R., & Krämer, N. (2006). Overview and Recent Advances in Partial Least Squares. In C. Saunders, M. Grobelnik, S. Gunn, & J. Shawe-Taylor (Eds.), *Subspace, Latent Structure and Feature Selection* (Vol. 3940, pp. 34–51). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved 2025-05-08, from http://link.springer.com/10.1007/11752790_2 (Series Title: Lecture Notes in Computer Science) doi: 10.1007/11752790_2

Ruan, X. (2020, March). Volatility-of-volatility and the cross-section of option returns. *Journal of Financial Markets*, *48*, 100492. Retrieved 2025-05-24, from https://linkinghub.elsevier.com/retrieve/

pii/S1386418118300818 doi: 10.1016/j.finmar.2019.03.002

Said, S. E., & Dickey, D. A. (1984). Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, *71*(3), 599–607. Retrieved 2025-05-21, from https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/71.3.599 doi: 10.1093/biomet/71.3.599

Shafaati, M., Chance, D. M., & Brooks, R. E. (2023). *The Cross-Section of Option Returns: Deriving Inferences in Sparse Models.* Retrieved 2025-05-25, from https://www.ssrn.com/abstract=4495089 doi: 10.2139/ssrn.4495089

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2011). Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *Journal of Statistical Software*, *39*(5). Retrieved 2025-05-07, from http://www.jstatsoft.org/v39/i05/ doi: 10.18637/jss.v039.i05

Tay, J. K., Narasimhan, B., & Hastie, T. (2023). Elastic Net Regularization Paths for All Generalized Linear Models. *Journal of Statistical Software*, *106*(1). Retrieved 2025-05-07, from https://www.jstatsoft.org/v106/i01/ doi: 10.18637/jss.v106.i01

Tian, M., & Wu, L. (2023, August). Limits of Arbitrage and Primary Risk-Taking in Derivative Securities. *The Review of Asset Pricing Studies*, *13*(3), 405–439. Retrieved 2025-05-24, from https://academic.oup.com/raps/article/13/3/405/7035950 doi: 10.1093/rapstu/raad003

Tibshirani, R. (1996, January). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), 267–288. Retrieved 2025-05-07, from https://academic.oup.com/jrsssb/article/58/1/267/7027929 doi: 10.1111/j.2517-6161.1996.tb02080.x

Timmermann, A. (2018, November). Forecasting Methods in Finance. *Annual Review of Financial Economics*, *10*(1), 449–479. Retrieved 2025-04-10, from https://www.annualreviews.org/doi/10.1146/annurev-financial-110217-022713 doi: 10.1146/annurev-financial-110217-022713

Utts, J. M. (1982, January). The rainbow test for lack of fit in regression. *Communications in Statistics - Theory and Methods*, *11*(24), 2801–2815. Retrieved 2025-05-20, from http://www.tandfonline.com/doi/abs/10.1080/03610928208828423 doi: 10.1080/03610928208828423

Vasquez, A. (2017, December). Equity Volatility Term Structures and the Cross Section of Option Returns. *Journal of Financial and Quantitative Analysis*, *52*(6), 2727–2754. Retrieved 2025-06-12, from https://www.cambridge.org/core/product/identifier/S002210901700076X/type/journal_article doi: 10.1017/S002210901700076X

Vasquez, A., & Xiao, X. (2024, April). Default Risk and Option Returns. *Management Science*, *70*(4), 2144–2167. Retrieved 2025-05-25, from https://pubsonline.informs.org/doi/10.1287/mnsc.2023.4796 doi: 10.1287/

mnsc.2023.4796

White, H. (2000, September). A Reality Check for Data Snooping. *Econometrica*, *68*(5), 1097–1126. Retrieved 2025-07-25, from http://doi.wiley.com/10.1111/1468-0262.00152 (Publisher: The Econometric Society) doi: 10.1111/1468-0262.00152

Wold, H. (1975). Soft Modelling by Latent Variables: The Non-Linear Iterative Partial Least Squares (NIPALS) Approach. *Journal of Applied Probability*, *12*(S1), 117–142. Retrieved 2025-05-07, from https://www.cambridge.org/core/product/identifier/S0021900200047604/type/journal_article doi: 10.1017/S0021900200047604

Wold, S., Sjöström, M., & Eriksson, L. (2001, October). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, *58*(2), 109–130. Retrieved 2025-06-09, from https://linkinghub.elsevier.com/retrieve/pii/S0169743901001551 doi: 10.1016/S0169-7439(01)00155-1

Yang, S., Aretz, K., Liu, H., & Zhang, Y. (2022, December). Consumption risks in option returns. *Journal of Empirical Finance*, *69*, 285–302. Retrieved 2025-05-25, from https://linkinghub.elsevier.com/retrieve/pii/S0927539822000883 doi: 10.1016/j.jempfin.2022.10.001

Yuan, J., Liu, D., Chen, C. R., & Hu, S. (2024, September). Option trading volume and the cross-section of option returns. *The North American Journal of Economics and Finance*, *74*, 102229. Retrieved 2025-05-25, from https://linkinghub.elsevier.com/retrieve/pii/S1062940824001542 doi: 10.1016/j.najef.2024.102229

Zhan, X. E., Han, B., Cao, J., & Tong, Q. (2022, February). Option Return Predictability. *The Review of Financial Studies*, *35*(3), 1394–1442. Retrieved 2025-05-24, from https://academic.oup.com/rfs/article/35/3/1394/6294944 doi: 10.1093/rfs/hhab067

Zou, H., & Hastie, T. (2005, April). Regularization and Variable Selection Via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *67*(2), 301–320. Retrieved 2025-05-07, from https://academic.oup.com/jrsssb/article/67/2/301/7109482 doi: 10.1111/j.1467-9868.2005.00503.x

A Related literature

A.1 Option factors

Find option factors discussed in recent literature below. The review includes, but is not limited, to many of the option factors used and described in Table B1.

Volatility and tail risk: Goyal and Saretto (2009) find that implied minus realized volatility predicts the expected returns of options in the cross-section. Constantinides, Jackwerth, and Savov (2013) show that adding crisis-related factors - particularly those capturing market price jumps, volatility jumps, and liquidity - alongside the market factor significantly reduces pricing errors. They highlight that non-normality of option returns can complicate factor construction. Cao and Han (2013) find that delta-hedged equity option return decreases monotonically with an increase in the idiosyncratic volatility of the underlying stock. Similarly, Hu and Jacobs (2020) show that the historical volatility of the underlying stock explains expected returns of options. The paper of Aretz, Lin, and Poon (2023) adds some nuance. It reveals that option returns vary in response to changes in volatility, depending on whether the changes are driven by systematic or idiosyncratic factors. Systematic volatility increases the expected returns of in-the-money and at-the-money calls, while decreasing the expected returns of out-of-the-money calls. In contrast, idiosyncratic volatility negatively affects expected call returns. Adding to this, Tian and Wu (2023) find that stochastic movements of volatility and jump risk of the underlying that may remove the delta-hedge are option factors. Also, they discover that delta-hedge does not fully remove the remaining risk because of limits to arbitrage. Hence, delta-hedging costs help explaining the expected returns. Duarte, Jones, and Wang (2024) emphasize the need to clean the microstructure noise when estimating the volatility risk premium on options. They confirm that volatility risk is negatively priced in single-stock options, just as it is for the S&P 500 index, especially for the heavy-volume segment of the market. Ruan (2020) shows that the stock's volatility-of-volatility is negatively related with its delta-hedged option returns. When looking at the distributional properties, (Bali & Murray, 2013) find that risk-neutral skewness predicts delta-hedged returns.

Liquidity: Christoffersen, Goyenko, Jacobs, and Karoui (2018) provide evidence that market makers in the equity options market hold large and risky net long positions, for which positive illiquidity premia compensate them. They prove this by calculating a risk-adjusted return spread for illiquid versus liquid equity options based on intraday effective spreads. Kanne, Korn, and Uhrig-Homburg (2023) find a similar effect of the underlying asset: the authors find that when the underlying stock is illiquid, options on it tend to yield higher expected returns if end users are net buyers. Muravyev (2016) shows that the inventory risk market makers in-

cur leads them to shade their quotes, reducing or increasing liquidity; Resulting imbalances in order flow then translate into higher (or lower) option prices until inventories are rebalanced. Yuan, Liu, Chen, and Hu (2024) find option trading volume negatively and significantly predicts the cross-section of delta-hedged option returns. This can be explained through idiosyncratic volatility and market capitalization.

Leverage and financial risk: Frazzini and Pedersen (2022) discover that options with higher embedded leverage alleviate investors leverage constraints more, resulting in lower returns. Vasquez and Xiao (2024) study the default risk of the underlying asset, measured by credit ratings or default probability. The authors find that delta-hedged option returns are negatively related. Zhan et al. (2022) find that expected option returns are negatively correlated with firm characteristics like stock price, profit margin, and overall profitability. They are positively correlated with high cash holdings, cash-flow volatility, new share issuance, total external financing, distress risk, and analyst forecast dispersion. The authors explore channels such as financing needs, risk exposures, and distress signaling to explain the factors. A more separate field from financial risk, Ho, Kagkadis, and Wang (2024) show that political risk predicts negative expected returns of options. Analyzing Brexit, the authors find that this is driven by the jump risk coming from political uncertainty.

Momentum and market sentiment: Jeon, Kan, and Li (2025) find that stock price autocorrelation determine expected option returns. Byun and Kim (2016) study the lottery characteristics of underlying stocks. The authors find that stocks with more lottery-like features - specifically, skewness - underperform stocks with the fewest lottery-like features. Since this effect is stronger during periods of high investor sentiment, increased optimism could raise the prices of these options.[17] In a related matter, Boulatov, Eisdorfer, Goyal, and Zhdanov (2022) show that investor attention to the underlying stock prices influence option prices. The authors demand pressure for options on stocks that receive less attention (e.g., stocks on mini-indices, or after stock splits). Ramachandran and Tayal (2021) analyze short sale constraints and find that they drive equity option returns of overpriced put options. While investors drive up the demand, dealers command a high premium as compensation for the increased market making risk. This dynamic increases the expected returns of those puts. Yang, Aretz, Liu, and Zhang (2022) make a connection between consumption risk and option prices. The authors observe a positive and negative effect of consumption growth and consumption volatility, respectively.

[17]The paper builds on the work of Boyer and Vorkink (2014), who discovered the relationship. The authors focus on the intermediary effect, demonstrating that options are overpriced due to limits on arbitrage. A recent working paper by Käfer (2025) discusses the implications of lottery characteristics for stocks in the pharmaceutical industry.

A.2 Two assumptions for factor timing with PCA

Haddad et al. (2020) factor-timing approach is based on two assumptions:

- Assumption 1: Asset returns are spanned by a small set of priced risk factors. This means the SDF is a linear function of these factors.
- Assumption 2: Markets exhibit no near-arbitrage opportunities. This means the factor covariance structure is well-behaved. These assumptions ensure that the largest PC of returns capture true priced factors rather than noise, making them predictable and usable for timing strategies.

B Data

B.1 General data structure

Figure B1 provides an overview of how the data is structured and from where I take the data.

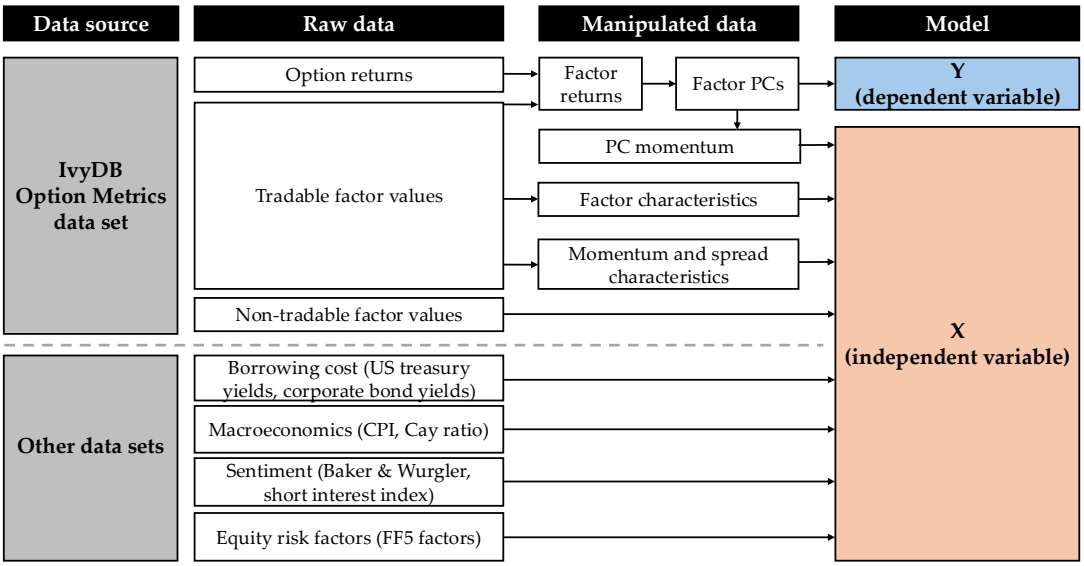


Figure B1: Simplified data pipeline. I use multiple data sources and apply various data manipulation steps to get the data for my model.

B.2 Factor portfolio returns

Table B1 presents the factors used. Note that I use the same factors as used by Käfer, Mörke, and Wiest (2025).

Variable name	Factor	Reference paper
embedlev	Embedded leverage	Frazzini and Pedersen (2022)
hc	Delta-hedging costs	Tian and Wu (2023)
vr	Volatility risk	Tian and Wu (2023)
jr	Historical jump risk	Tian and Wu (2023)
vov	Volatility of implied volatility	Ruan (2020)
optspread	Option illiquidity	Christoffersen, Goyenko, et al. (2018)
hvol	Historical stock volatility	Hu and Jacobs (2020)
sysvol	Systematic volatility	Aretz et al. (2023)
ivterm	Implied ATM volatility term structure	Vasquez (2017)
ac	Stock return autocorrelation	Jeon et al. (2025)
max10	Avg. of 10 highest past returns	Byun and Kim (2016)
defrisk	Default risk	Vasquez and Xiao (2024)
iskew	Idiosyncratic skewness	Byun and Kim (2016)
tskew	Total skewness	Byun and Kim (2016)
ivol	Idiosyncratic volatility	Cao and Han (2013)
ivrv	Implied minus realized volatility	Goyal and Saretto (2009)
amihud	Stock illiquidity	Zhan et al. (2022); Kanne et al. (2023)
rsi	Short interest	Ramachandran and Tayal (2021)
issue_1y	1-year new stock issues	Zhan et al. (2022)
issue_5y	5-year new stock issues	Zhan et al. (2022)
disp	Analyst dispersion	Zhan et al. (2022)
zscore	Altman Z-score	Zhan et al. (2022)
cash_at	Cash-to-assets ratio	Zhan et al. (2022)
ocfq_saleq_std	Cash flow volatility	Zhan et al. (2022)
ope_be	Operating profits / book equity	Zhan et al. (2022)
ebit_sale	Profit margin	Zhan et al. (2022)
netis_at	Net total issuance	Zhan et al. (2022)
log_price	Stock price	Boulatov et al. (2022)

Table B1: 28 option factors and their sources

When we look at the factor portfolio returns, we see that most factors have significant positive returns. Although they can be negative at times, most median values are also positive.

The p-values of the Jarque and Bera (1980) test indicate that the factor returns do not follow a normal distribution. Note that, while the factor returns are leptokurtotic, they are not negatively skewed in comparison to monthly equity returns (that have a fat left tail). While delta-hedged option returns are positively skewed (Bali et al., 2023), there does not seem to be a consistent pattern in their factor returns.

Factor	Mean	SD	Min	Q1	Median	Q3	Max	Skew	Kurt	JB
log_price	0.009***	0.015	-0.033	0.002	0.009	0.016	0.078	0.507	2.801	0.000
netis_at	0.005***	0.013	-0.074	-0.001	0.005	0.012	0.073	-0.167	6.496	0.000
ebit_sale	0.008***	0.013	-0.053	0.000	0.009	0.015	0.071	-0.150	2.959	0.000
ope_be	0.007***	0.013	-0.048	0.000	0.008	0.014	0.071	-0.213	3.211	0.000
ocfq_saleq_std	0.008***	0.012	-0.036	0.002	0.009	0.016	0.047	-0.397	0.788	0.000
cash_at	0.008***	0.014	-0.057	0.000	0.008	0.015	0.084	0.349	4.850	0.000
disp	0.003***	0.010	-0.042	-0.002	0.004	0.009	0.040	-0.359	2.652	0.000
zscore	0.002***	0.012	-0.027	-0.005	0.001	0.009	0.066	1.358	4.470	0.000
issue_5y	0.005***	0.011	-0.034	-0.001	0.006	0.012	0.061	-0.006	2.094	0.000
issue_1y	0.004***	0.012	-0.041	-0.004	0.004	0.010	0.063	0.818	3.851	0.000
rsi	0.003***	0.010	-0.037	-0.002	0.004	0.009	0.031	-0.401	1.444	0.000
amihud	0.005***	0.013	-0.031	-0.003	0.004	0.012	0.048	0.160	0.847	0.003
ivrv	0.023***	0.020	-0.023	0.011	0.019	0.032	0.100	0.945	1.239	0.000
ivol	0.008***	0.015	-0.075	0.001	0.010	0.017	0.066	-0.852	4.724	0.000
tskew	0.002***	0.008	-0.022	-0.003	0.002	0.006	0.045	0.532	2.533	0.000
iskew	0.002***	0.008	-0.029	-0.003	0.002	0.006	0.030	0.142	1.231	0.000
defrisk	0.002**	0.015	-0.086	-0.004	0.003	0.010	0.049	-0.998	5.655	0.000
max10	0.006***	0.019	-0.083	-0.003	0.009	0.016	0.072	-1.125	4.506	0.000
ac	0.001	0.008	-0.036	-0.004	0.001	0.005	0.029	-0.144	1.533	0.000
ivterm	0.009***	0.015	-0.043	0.001	0.009	0.017	0.061	0.047	1.436	0.000
sysvol	0.001	0.018	-0.044	-0.010	-0.002	0.008	0.090	1.728	5.652	0.000
hvol	0.007***	0.017	-0.079	-0.001	0.009	0.017	0.072	-0.847	3.925	0.000
optspread	0.002***	0.010	-0.039	-0.004	0.002	0.007	0.032	0.049	1.812	0.000
vov	0.004***	0.011	-0.041	-0.002	0.003	0.009	0.052	0.440	3.471	0.000
jr	0.008***	0.010	-0.013	0.001	0.007	0.012	0.056	1.204	3.373	0.000
vr	0.010***	0.014	-0.039	0.002	0.009	0.018	0.077	0.370	2.552	0.000
hc	0.007***	0.013	-0.026	-0.001	0.007	0.015	0.066	0.474	1.146	0.000
embedlev	0.018***	0.017	-0.037	0.010	0.018	0.027	0.114	0.572	3.742	0.000

Table B2: Summary statistics for factor returns. Includes mean, standard deviation (SD), quantiles (Q1, Q3), range, skewness, excess kurtosis, and p-value of the Jarque and Bera (1980) test for normality. All values rounded to 3 d.p. Means annotated with significance, based on t-statistic: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

B.3 Predictors

Non-traded factors. I use non-traded factors from Käfer, Moerke, et al. (2025). Here is a list taken from their paper:

1. Intermediary capital nontraded risk (CPTL): The intermediary capital nontraded risk factor of He, Kelly, and Manela (2017). The data is downloaded from Zhiguo He's website at https://voices.uchicago.edu/zhiguohe/.

2. Economic policy uncertainty (EPU): The first difference in the economic policy uncertainty index. The data is taken from FRED.

3. Macroeconomic uncertainty (UNC): The first difference in the macroeconomic uncertainty index lagged by one month to align the forecast to the returns observed in month t (Jurado, Ludvigson, & Ng, 2015). The data is taken from Sydney Ludvigson's website at https://www.sydneyludvigson.com/.

4. Financial economic uncertainty (UNCf): The first difference in the financial economic uncertainty index lagged by one month to align the forecast to the returns observed in month t. The data is taken from Sydney Ludvigson's website at https://www.sydneyludvigson.com/.

5. Real economic uncertainty (UNCr): The first difference in the real economic uncertainty index lagged by one month to align the forecast to the returns observed in month t. Data is taken from Sydney Ludvigson's website at https://www.sydneyludvigson.com/.

6. Volatility risk, VIX (VIX): The first difference in the CBOE VIX index. The data is downloaded from https://www.cboe.com/tradable products/vix/vix historical data/.

7. Volatility-of-volatility risk, VIXVOL (VIXVOL): A range-based measure of the volatility of aggregate volatility based on daily readings of the VIX index. The construction follows Agarwal, Arisoy, and Naik (2017).

8. Market tail risk (SKEW): The first difference in the CBOE SKEW index which estimates market tail risk. The data is downloaded from https://www.cboe.com/us/indices/c

9. SKEW term structure (SKEWTS): The first difference in the CBOE SKEW term structure. It is the difference between 182d and 30d SKEW. Data downloaded from https://www.cboe.com/us/indices/dashboard/skew/.

10. Correlation risk (ICRC): The payoff of monthly S&P 500 correlation swaps which is the difference between the implied and realized correlation (Buraschi, Kosowski, & Trojani, 2014).

11. Aggregate liquidity risk (LIQ): The aggregate liquidity innovation factor of Pastor and Stambaught (2003). The data is downloaded from Robert Stambaugh's website at https://finance.wharton.upenn.edu/ stambaug/ .

12. Sentiment (SENT): The sentiment index of Baker and Wurgler (2006) orthogonalized with respect to macroeconomic variables. The data is obtained from Jeffrey Wurgler's website at https://pages.stern.nyu.edu/ jwurgler/.

13. Interest rate term structure (TERM): The slope of the term structure of interest rates. TERM is calculated as the difference between U.S. Treasury Securities

at 10-year constant maturity and 3-month Treasury bill secondary market rates. The data is obtained from FRED.

14. Default spread (DEF): The default spread defined as the yield difference between Moody's AAA and BAA corporate bond yields. The data is taken from FRED.
15. Financial Stress (STLFSI): The St. Louis Fed Financial Stress Index. Data is obtained from FRED.

Following Horenstein et al. (2023), I control for general option market risks by including *ew_ret*, an equal-weighted portfolio of all 280 decile portfolios used in the factor construction.

C Estimation procedures

C.1 PCA

As mentioned above, PCA transforms the data in a linear way such that each resulting principle component series is uncorrelated with the others. Take the vector of factor returns x_t with length n. Then linearly transform it to have the highest variance possible $q_{t,1}^\top x_t = PC_{t,1}$. Then, create another linear combination with maximum variance, that is also uncorrelated to $q_{1,t,1}x_{1,t}, \ldots, q_{n,t,1}x_{n,t}$: $q_{t,2}^\top x_t = PC_{t,2}$. This is continued until $PC_{t,n}$. In matrix algebra, we can write

$$PC = XQ, \quad \text{with } PC \in \mathbb{R}^{T\times N}, X \in \mathbb{R}^{T\times N}, Q \in \mathbb{R}^{N\times N}, \tag{13}$$

where Q is the matrix of eigenvectors of the covariance matrix Σ of X. See Jolliffe (2002) for a detailed discussion of derivation and properties of PCA.

C.2 Models

Benchmark. I use an ARMA model to test whether simple autocorrelation can better explain movements in the PCs compared to factor weights or other predictors. The ARMA(p, q) model is defined as:

$$\hat{y}_{t+1}^{\text{ARMA}} = \hat{a}_0 + \sum_{i=1}^{p} \hat{a}_i y_{t+1-i} + \sum_{j=1}^{q} \hat{\beta}_j \hat{\varepsilon}_{t+1-j}, \tag{14}$$

where p and q are the autoregressive and moving average lag lengths, respectively. The coefficients $\hat{a}_i$ and $\hat{\beta}_j$ are estimated via maximum likelihood, assuming Gaus-

sian errors. The lag orders are selected using the corrected Akaike Information Criterion (Hyndman & Khandakar, 2008).

I test the unit root behaviour of the PCs using the ADF test. I find that all PCs behaved like stationary time series, indicating that it is not necessary to integrate the series for the use of ARMA models (Said & Dickey, 1984).

Linear models. The Lasso, Ridge, and Elastic Net (EN) regressions solve the problem

$$\hat{\beta}^{\text{EN}} = \min_{\beta} \sum_{t=1}^{T} (y_t - X_t^\top \beta)^2 + \lambda \left[\underbrace{\sum_{j=1}^{p} \alpha |\beta_j|}_{\text{L1 penalty (Lasso)}} + \underbrace{\sum_{j=1}^{p} (1-\alpha)\beta_j^2}_{\text{L2 penalty (Ridge)}} \right], \tag{15}$$

where X_t includes a column of ones to account for a constant, y_t is the respective PC value, and p is the number of parameters. Depending on what model I use, I adapt the values of α: it equals one for the Lasso, zero for the Ridge, and something in between for the EN. In case of the EN regression, I choose α by cross-validation using 0.1-steps between zero and one. The λ parameter is chosen by cross validation as well. It is the regulization strength (Tay, Narasimhan, & Hastie, 2023).

When $\alpha = 0$, the Elastic Net objective reduces to Ridge regression, which applies an L2 penalty. Through adding the sum of squares of the coefficient estimates, the penalty shrinks coefficients toward zero. This allows the estimates to be more general, especially in case of high dimensionality with limited number of observations (Hoerl & Kennard, 1970) .

When $\alpha = 1$, the Elastic Net reduces to Lasso regression, which applies a pure L1 penalty. Unlike Ridge regression, the L1 penalty has a sharp peak at zero, which encourages sparsity by shrinking some coefficients exactly to zero. This leads to automatic variable selection, as irrelevant predictors are effectively excluded from the model. In contrast, the squared L2 penalty used in Ridge regression does not penalise coefficients close to zero too much and keeps all coefficients nonzero, even if small. Lasso is useful in high-dimensional settings where only a few predictors are expected to be relevant, and interpretability is important (Tibshirani, 1996). However, Lasso tends to struggle when predictors are highly correlated — it selects one and ignores others.

The Elastic Net combines Lasso and Ridge penalties, requiring the estimation of an additional mixing parameter. It is particularly useful when predictors are correlated, since Lasso tends to select only one predictor and discard the rest, whereas

the Elastic Net allows for grouped selection and shrinkage (Zou & Hastie, 2005; Tay et al., 2023).

As another method, I use PLS. It was introduced by H. Wold (1975). I use the kernel method to solve the method, introduced by Dayal and MacGregor (1997). In the following, instead of going into the mathematical details, I focus on the main functioning of the model.

PLS regression is a dimension-reduction technique which constructs latent variables (components) as linear combinations of the initial predictors such that their covariance with the response variable is optimized. Contrary to traditional OLS, where predictors would induce instability or singularity when it becomes or in the worst-case scenario of too many regressors p versus too few observations n (where $p > n$), PLS merely overcomes those issues because it projects data to a lower, smaller space that records most relevant variation of predictor matrix X with the same predictive power in relation to the response Y. Therefore, PLS is particularly valuable in high-dimensional settings, such as macroeconomic forecasting or chemometrics, where predictors may be highly correlated and standard regression methods do not work (S. Wold et al., 2001).

Nonlinear models. Random Forest is a ML method that builds many decision trees and averages all the predictions to make a final prediction. Each tree is trained using a bootstrapped subset of the data, considering only a random subset of features at each split. This prevents overfitting and makes the method robust to noise (Gu et al., 2020). The original algorithm was introduced by Breiman (2001).

XGBoost is a scalable implementation of gradient boosting decision trees. Trees are built sequentially, with each one trained to correct the residual errors of the previous ensemble. It uses gradient descent to minimize a differentiable loss function (often squared error or log loss) and includes explicit regularization (e.g., L1/ L2 penalties, tree pruning, shrinkage) to prevent overfitting. Gradient boosting was introduced by J. H. Friedman (2001); XGBoost was developed later by T. Chen and Guestrin (2016).

Dart (Dropout Additive Regression Trees), a variant of XGBoost, includes a dropout mechanism during training. Some trees are randomly omitted when a new tree is added to the ensemble. This reduces the risk of overfitting because non-generalizable features are excluded, adding another layer of regularisation. Dart outperforms standard boosting methods in settings where model complexity or redundancy is high. It combines the strengths of boosting with the concept of dropout in neural networks (Gu et al., 2020). The Dart model was introduced by Rashmi and Gilad-Bachrach (2015).

C.3 R-packages used for ML models

I use the following R-packages in my thesis for model estimation:

- Lasso, Ridge, and Elastic Net: I use the `glmnet` package (J. Friedman, Hastie, & Tibshirani, 2010; Simon, Friedman, Hastie, & Tibshirani, 2011; Tay et al., 2023).
- Partial Least Squares: I use the `pls` package (Mevik & Wehrens, 2007).
- ARIMA: I use the `forecast` package (Hyndman & Khandakar, 2008) .
- Random Forest, XGBoost, and Dart: I use the `xgboost` package (T. Chen & Guestrin, 2016).

C.4 Hyperparameter space for parameter optimization

The hyperparameter optimization is based on Bali et al. (2023). However, for performance improvement, I modify and reduce the parameter grid minorly. Tables C1, C2, and C3 show the used hyperparameter space.

Random Forest hyperparameter	**Values**
Max depth per tree	$\{3, 6\}$
Row subsample fraction	$\{0.7, 0.9\}$
Feature subsample fraction (per tree)	$\{0.7, 0.9\}$
Number of trees	$\{50, 100, 200\}$

Table C1: Random Forest hyperparameter space.

XGBoost hyperparameter	**Values**
Learning rate (η)	$\{0.03, 0.10\}$
Max depth per tree	$\{3, 6\}$
Row subsample fraction	0.8
Feature subsample fraction (per tree)	$\{0.7, 0.9\}$
Minimum child weight	$\{1, 5\}$
L2 regularization (λ)	$\{1, 5\}$
L1 regularization (α)	$\{0, 0.5\}$

Table C2: XGBoost hyperparameter space.

Dart hyperparameter	Values
Learning rate (η)	$\{0.03, 0.10\}$
Max depth per tree	$\{3, 6\}$
Row subsample fraction	$\{0.7, 0.9\}$
Feature subsample fraction (per tree)	$\{0.7, 0.9\}$
Dropout rate (`rate_drop`)	$\{0.05, 0.10, 0.15\}$
Dropout skip probability (`skip_drop`)	0.25
Sample type	`uniform` (fixed)
Normalization type	`tree` (fixed)
Gamma	0 (fixed)

Table C3: Dart hyperparameter space.

D Further results

	real_return	lasso	ridge	en	pls	linear_ens	hist	arima	nonlinear_ens	XGBoost	Dart	rand_forest
real_return		0.77	0.96	0.78	0.61	0.83	0.99	0.91	0.94	0.87	0.88	0.99
lasso			0.84	0.98	0.81	0.96	0.76	0.75	0.86	0.84	0.86	0.82
ridge				0.85	0.74	0.91	0.96	0.92	0.95	0.90	0.92	0.97
en					0.81	0.97	0.77	0.75	0.87	0.85	0.87	0.83
pls						0.90	0.59	0.63	0.70	0.69	0.72	0.65
linear_ens							0.82	0.82	0.90	0.87	0.90	0.87
hist								0.91	0.93	0.86	0.88	0.98
arima									0.88	0.82	0.85	0.91
nonlinear_ens										0.98	0.99	0.97
XGBoost											0.97	0.92
Dart												0.93
rand_forest												

Table D1: Model forecasting correlation for PC1

	real_return	lasso	ridge	en	pls	linear_ens	hist	arima	nonlinear_ens	XGBoost	Dart	rand_forest
real_return		0.89	0.92	0.89	0.70	0.89	0.91	0.88	0.95	0.90	0.92	0.99
lasso			0.94	0.96	0.78	0.97	0.88	0.88	0.92	0.90	0.90	0.91
ridge				0.95	0.84	0.98	0.94	0.93	0.95	0.93	0.94	0.94
en					0.80	0.97	0.88	0.89	0.92	0.91	0.91	0.91
pls						0.90	0.73	0.82	0.78	0.77	0.79	0.74
linear_ens							0.90	0.92	0.93	0.92	0.93	0.92
hist								0.85	0.92	0.89	0.89	0.92
arima									0.91	0.88	0.89	0.90
nonlinear_ens										0.99	0.99	0.98
XGBoost											0.97	0.94
Dart												0.95
rand_forest												

Table D2: Model forecasting correlation for PC2

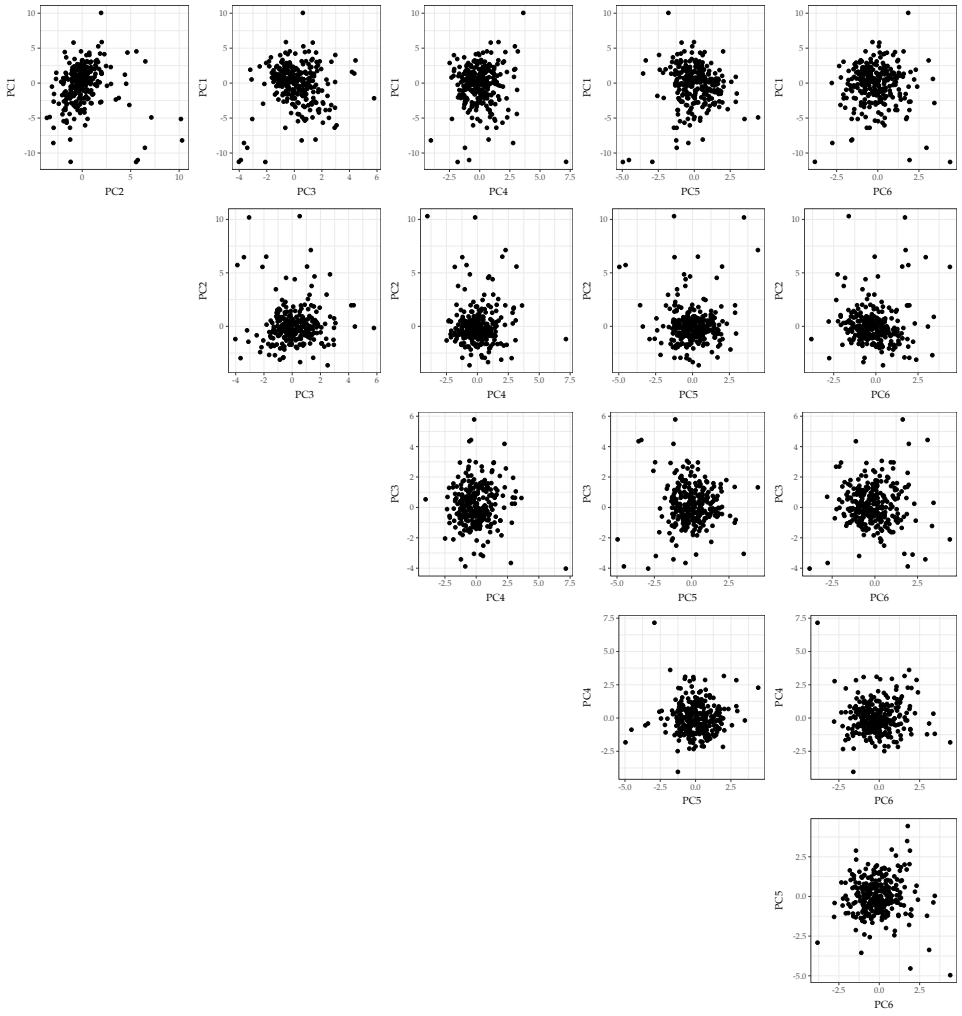


Figure D1: Scatter plot for each PC combination. By design, PCs do not have any correlation between each other. We cannot observe any visible patterns from these graphs.

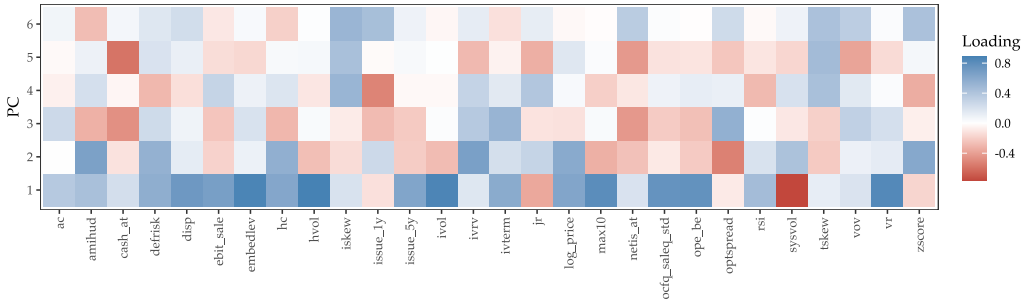


Figure D2: PC loading heatmap. We observe that for more relevant PCs, the loadings are stronger.

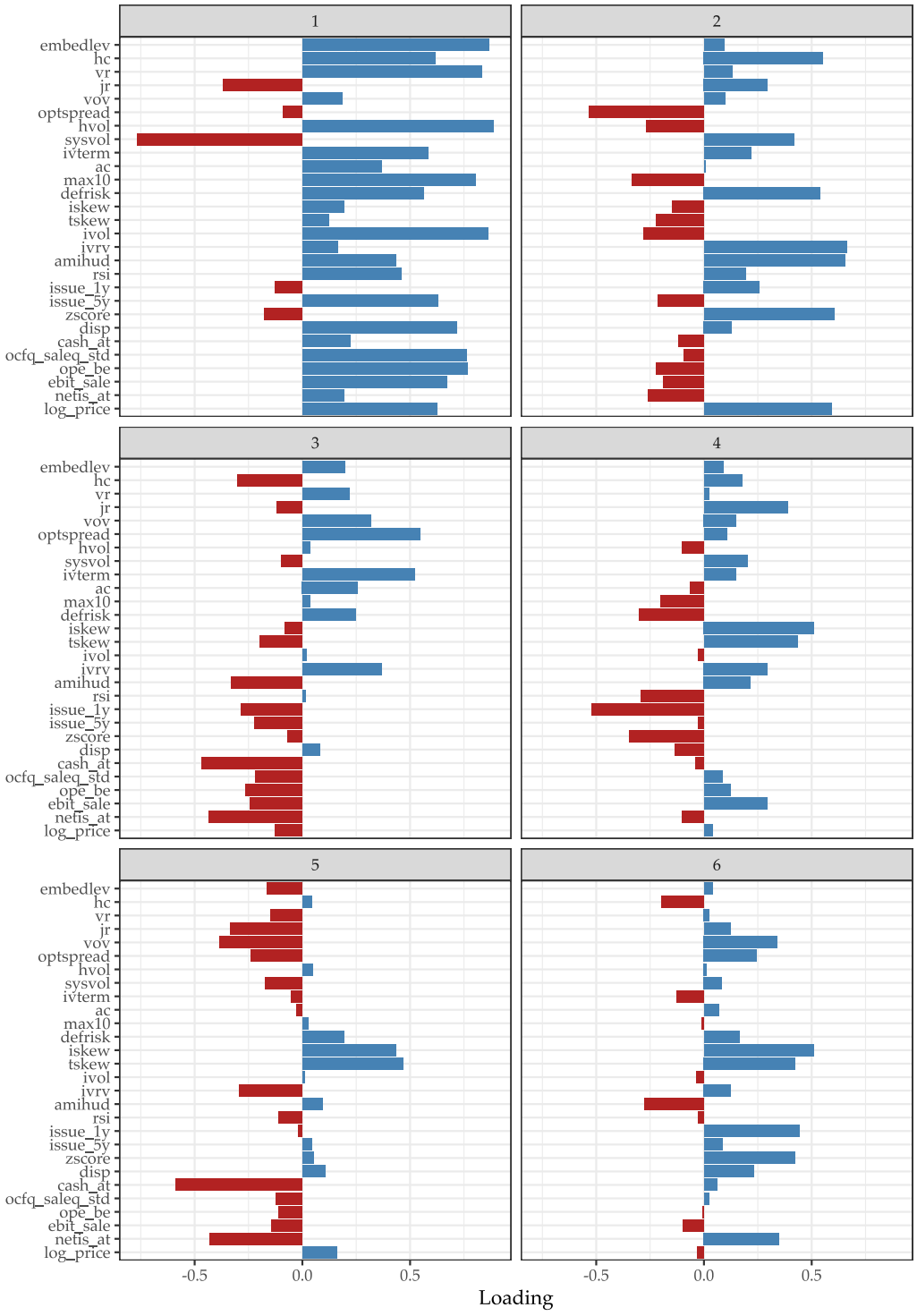


Figure D3: PC loading bar chart, representing Table 5 visually. Each graph represents the loadings for a PC.

	real_return	lasso	ridge	en	pls	linear_ens	hist	arima	nonlinear_ens	XGBoost	Dart	rand_forest
real_return		0.90	0.93	0.90	0.68	0.89	0.97	0.93	0.94	0.89	0.91	0.99
lasso			0.95	0.97	0.77	0.97	0.93	0.92	0.94	0.93	0.92	0.93
ridge				0.97	0.81	0.98	0.97	0.95	0.96	0.94	0.94	0.96
en					0.80	0.98	0.93	0.93	0.95	0.93	0.93	0.93
pls						0.89	0.73	0.75	0.77	0.77	0.77	0.73
linear_ens							0.93	0.93	0.95	0.93	0.93	0.92
hist								0.95	0.96	0.92	0.93	0.98
arima									0.94	0.91	0.93	0.95
nonlinear_ens										0.98	0.99	0.97
XGBoost											0.97	0.93
Dart												0.94
rand_forest												

Table D3: Model forecasting correlation for PC3

	real_return	lasso	ridge	en	pls	linear_ens	hist	arima	nonlinear_ens	XGBoost	Dart	rand_forest
real_return		0.83	0.94	0.82	0.79	0.89	0.95	0.93	0.95	0.91	0.92	0.99
lasso			0.90	0.94	0.81	0.96	0.87	0.83	0.85	0.82	0.84	0.84
ridge				0.89	0.85	0.96	0.97	0.94	0.96	0.93	0.95	0.96
en					0.84	0.96	0.85	0.82	0.85	0.83	0.85	0.84
pls						0.92	0.81	0.80	0.84	0.84	0.82	0.81
linear_ens							0.92	0.89	0.92	0.90	0.91	0.91
hist								0.95	0.95	0.91	0.93	0.96
arima									0.93	0.89	0.90	0.94
nonlinear_ens										0.98	0.99	0.98
XGBoost											0.96	0.94
Dart												0.95
rand_forest												

Table D4: Model forecasting correlation for PC4

	real_return	lasso	ridge	en	pls	linear_ens	hist	arima	nonlinear_ens	XGBoost	Dart	rand_forest
real_return		0.89	0.95	0.89	0.67	0.90	0.95	0.94	0.95	0.91	0.91	0.99
lasso			0.94	0.97	0.69	0.96	0.93	0.91	0.92	0.91	0.89	0.91
ridge				0.95	0.71	0.96	0.98	0.97	0.96	0.94	0.94	0.96
en					0.71	0.97	0.92	0.92	0.93	0.91	0.90	0.91
pls						0.84	0.68	0.69	0.69	0.67	0.68	0.68
linear_ens							0.93	0.93	0.93	0.92	0.91	0.92
hist								0.97	0.96	0.93	0.94	0.96
arima									0.95	0.92	0.93	0.96
nonlinear_ens										0.99	0.99	0.98
XGBoost											0.97	0.94
Dart												0.95
rand_forest												

Table D5: Model forecasting correlation for PC5

	real_return	lasso	ridge	en	pls	linear_ens	hist	arima	nonlinear_ens	XGBoost	Dart	rand_forest
real_return		0.88	0.92	0.88	0.82	0.90	0.94	0.91	0.95	0.91	0.92	0.99
lasso			0.95	0.98	0.86	0.98	0.91	0.91	0.94	0.92	0.93	0.91
ridge				0.95	0.90	0.98	0.97	0.95	0.97	0.95	0.96	0.95
en					0.87	0.98	0.90	0.90	0.93	0.92	0.93	0.91
pls						0.94	0.86	0.86	0.87	0.86	0.86	0.85
linear_ens							0.94	0.93	0.96	0.94	0.95	0.93
hist								0.94	0.96	0.93	0.94	0.96
arima									0.93	0.91	0.92	0.93
nonlinear_ens										0.99	0.99	0.98
XGBoost											0.97	0.94
Dart												0.96
rand_forest												

Table D6: Model forecasting correlation for PC6

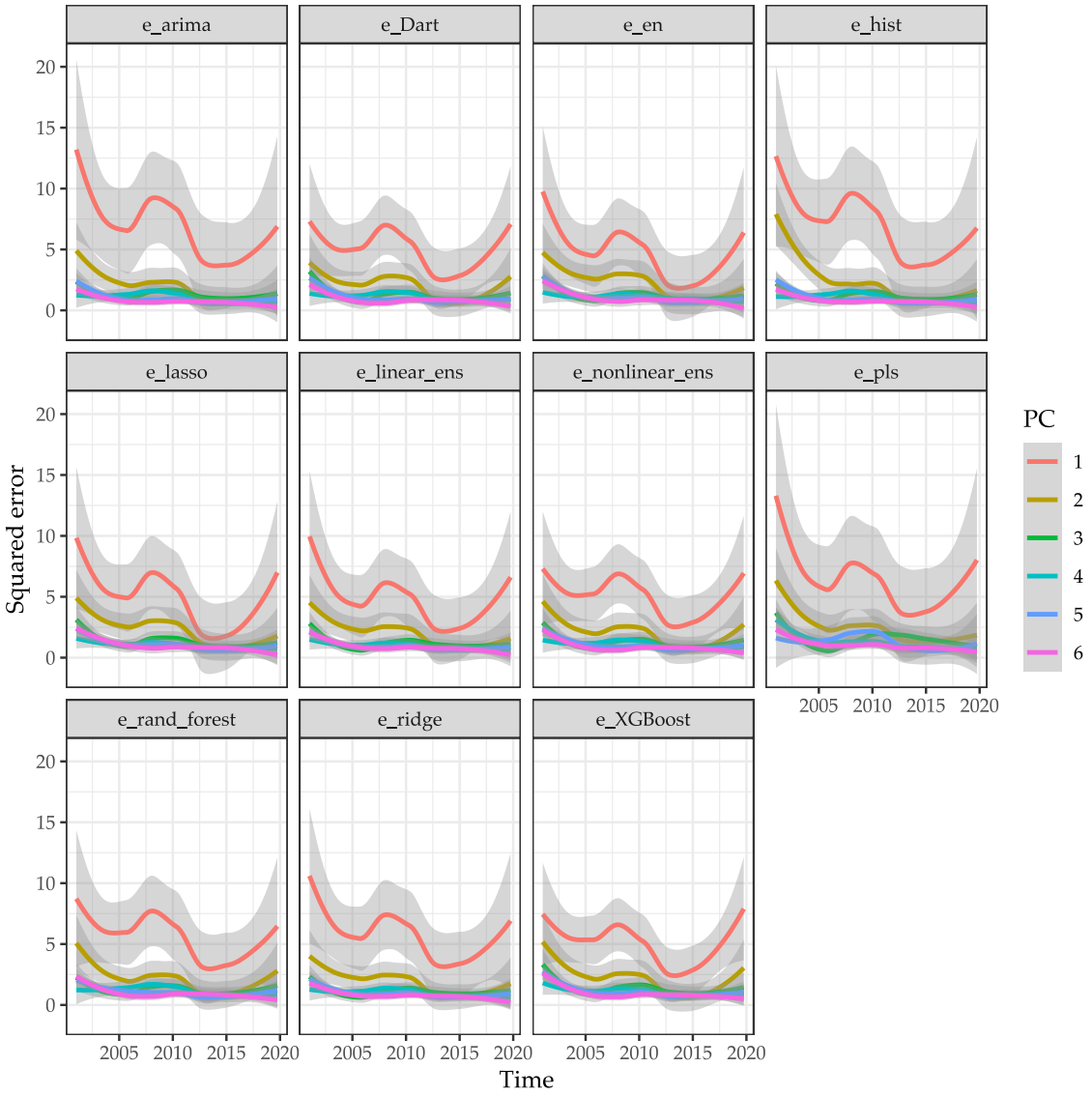


Figure D4: Error values over time. Smoothed by Cleveland and Devlin (1988)

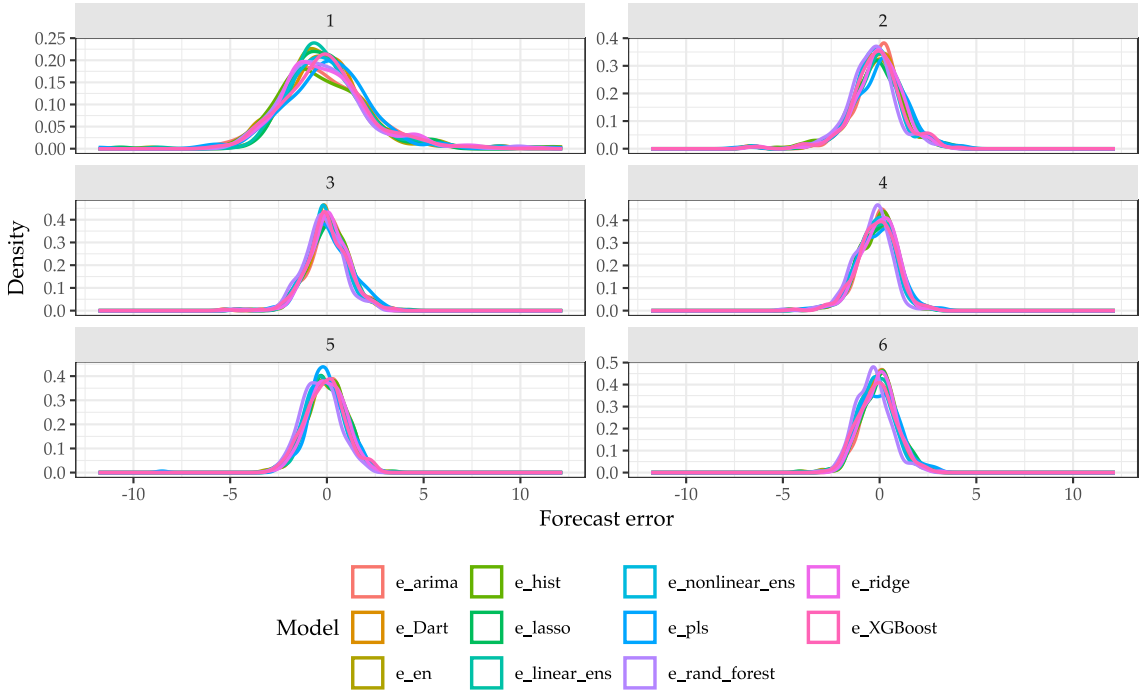


Figure D5: Error densities for each model and PC.

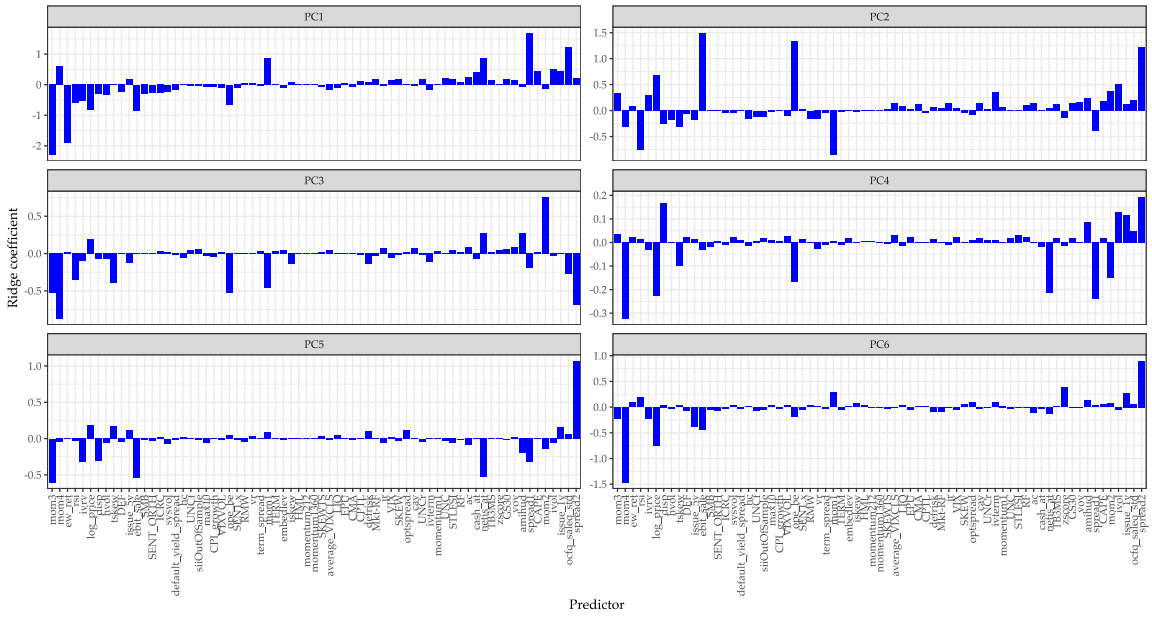


Figure D6: Ridge coefficients.

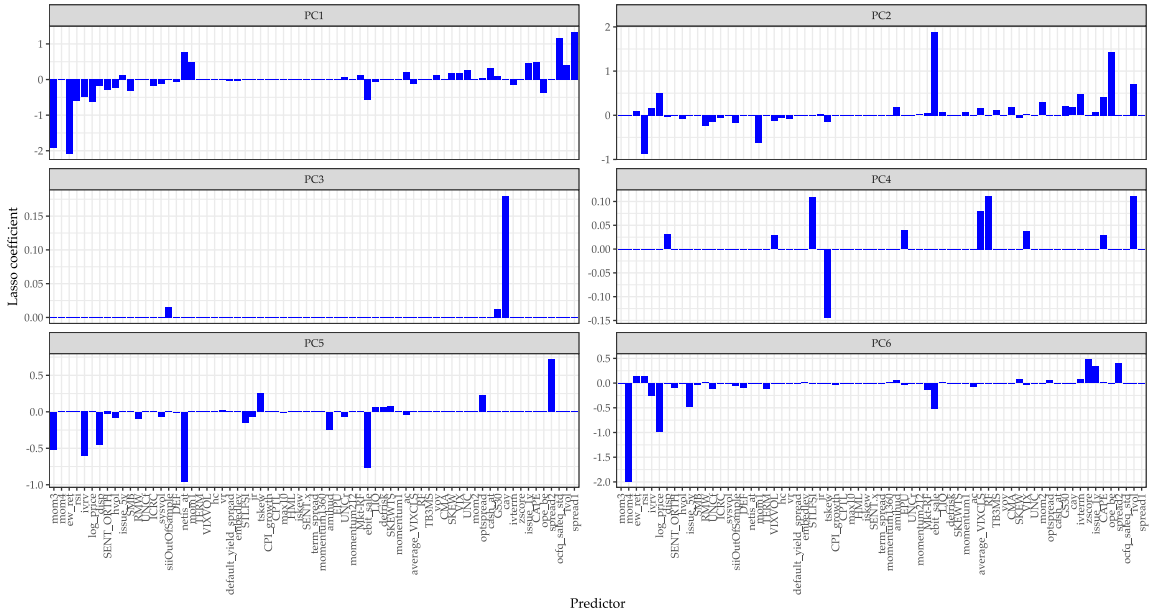


Figure D7: Lasso coefficients.

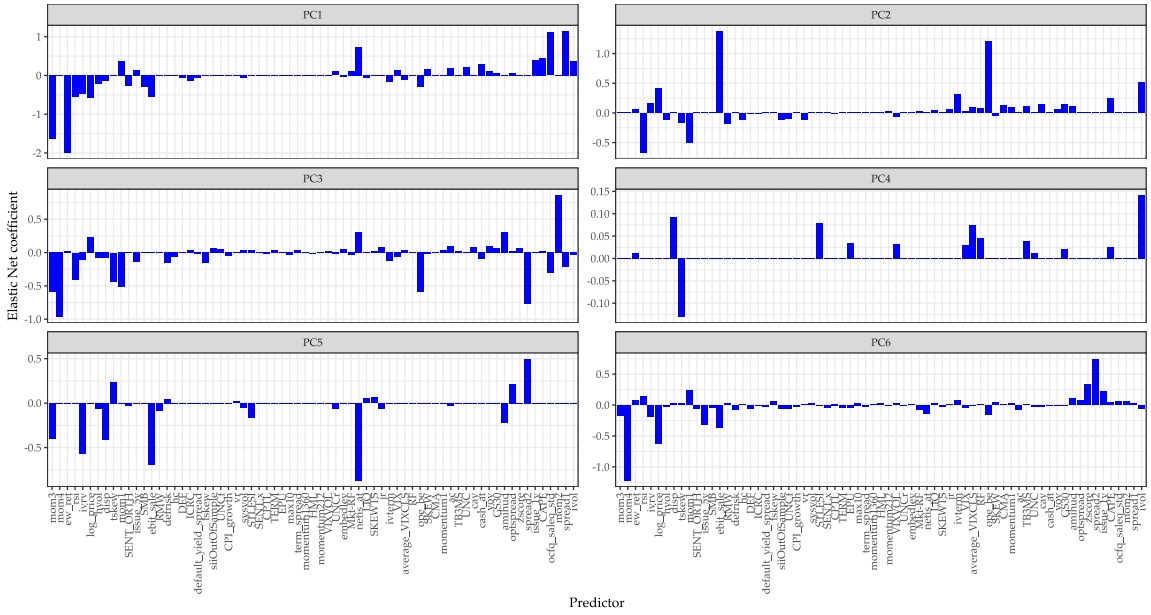


Figure D8: Elastic Net coefficients.

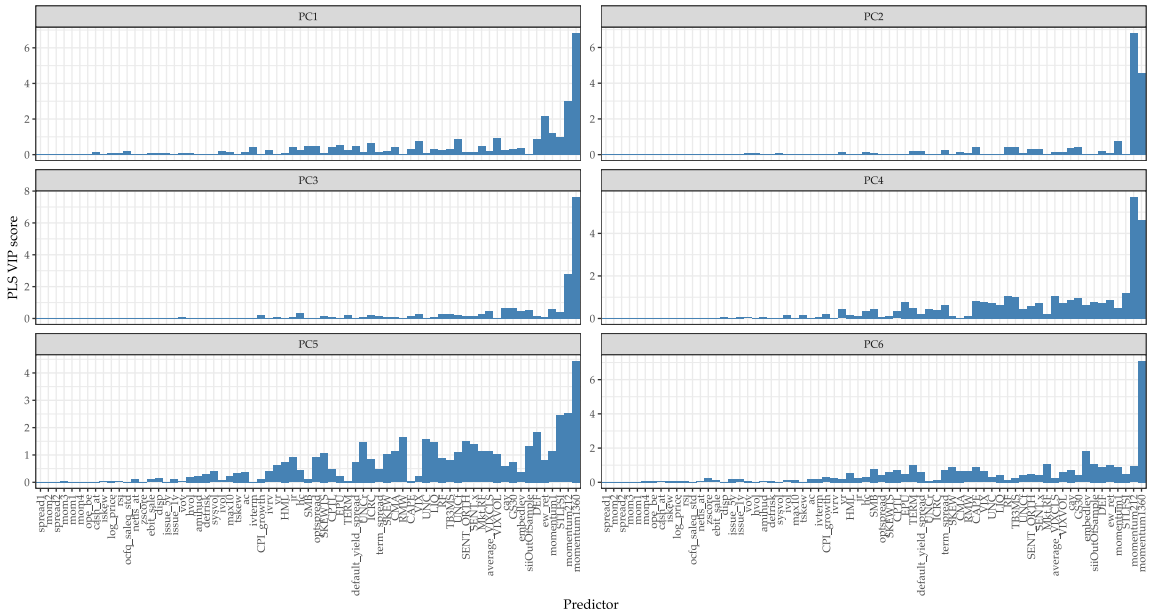


Figure D9: PLS Value Importance.

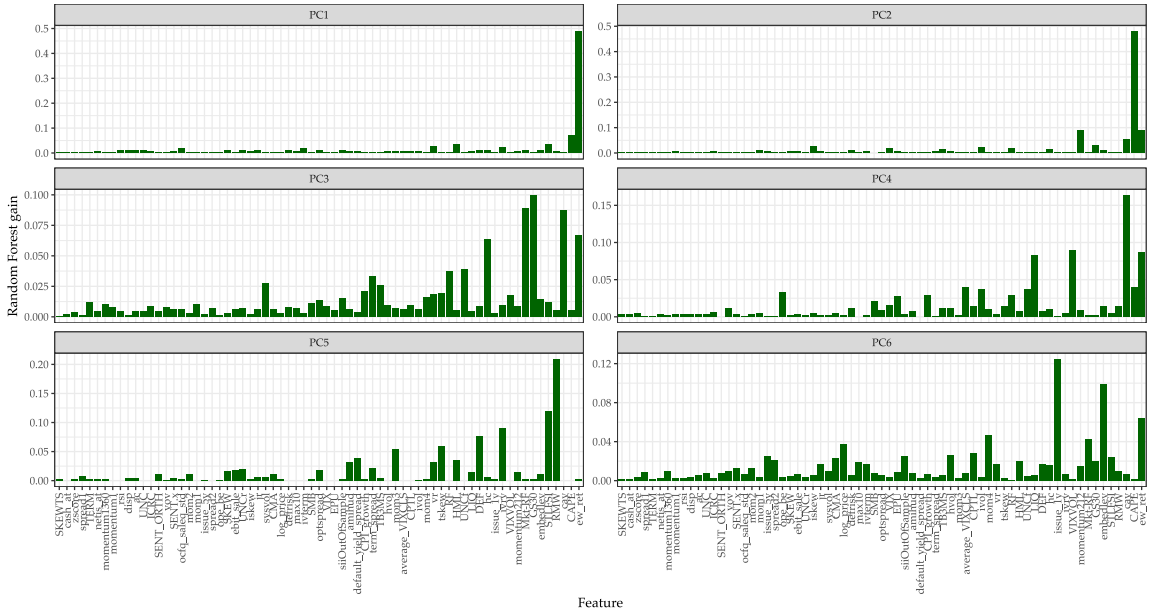


Figure D10: Random Forest feature importance by gain. This shows how much a feature improves the model's accuracy when it is used to split the data - on average, across all trees in the model.

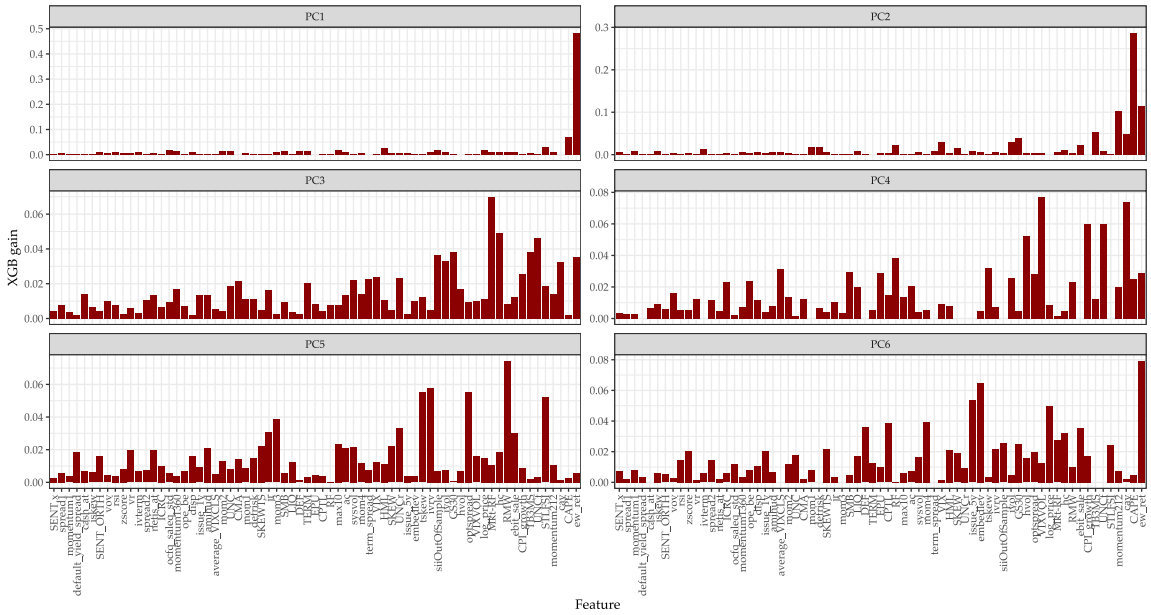


Figure D11: XGBoost feature importance by gain. This shows how much a feature improves the model's accuracy when it is used to split the data - on average, across all trees in the model.

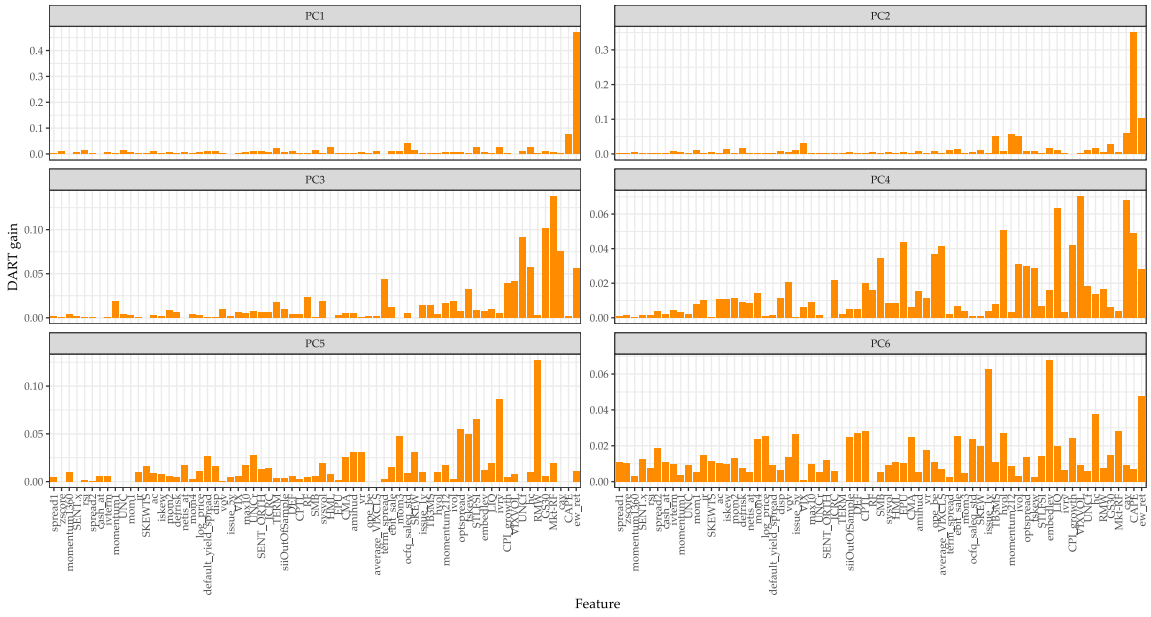


Figure D12: Dart feature importance by gain. This shows how much a feature improves the model's accuracy when it is used to split the data - on average, across all trees in the model.

Model	Return	Total return	SD	Sortino Ratio	Hit Ratio	Sharpe Ratio	SR CI low	SR CI high
arima	0.10	2.21	0.11	1.61	0.67	0.96	0.57	1.45
equal_weight	0.04	1.11	0.02	3.96	0.85	2.71	2.15	3.40
hist	0.06	1.33	0.11	0.85	0.62	0.54	0.08	1.02
linear_ens	0.21	4.26	0.12	3.34	0.77	1.78	1.36	2.22
nonlinear_ens	0.17	3.50	0.12	3.30	0.74	1.49	1.13	1.86

Table D7: Model performance overview using the RiskMetrics approach. *Return*: annualized excess return; *Total return*: the return over the entire period; *SD*: annualized standard deviation of returns; *Sortino Ratio*: downside-risk adjusted return ($\bar{r}/SD(r^-)$, where $SD(r^-)$ is the standard deviation of negative returns); *Hit Ratio*: fraction of periods with positive returns; *Sharpe Ratio*: risk adjusted returns ($r - \bar{r}_{rf}/SD(r)$); *SR CI low* and *SR CI high*: 95% confidence interval. Confidence bands are created by bootstrapping, following Ledoit and Wolf (2008): I take 1000 bootstrap samples from the return series (with replacement) and from these simulations retrieve the confidence bands of the Sharpe Ratio. For the risk-free rate r_{rf}, I use the 3-Month Treasury Bill Secondary Market Rate (Discount Basis).

Model	Return	Total return	SD	Sortino Ratio	Hit Ratio	Sharpe Ratio	SR CI low	SR CI high
arima	0.12	2.55	0.14	1.43	0.68	0.86	0.44	1.31
equal_weight	0.04	1.11	0.02	3.96	0.85	2.71	2.15	3.40
hist	0.03	0.92	0.13	0.37	0.60	0.26	-0.18	0.76
linear_ens	0.25	4.88	0.15	3.21	0.73	1.61	1.20	2.07
nonlinear_ens	0.21	4.12	0.16	2.59	0.73	1.25	0.93	1.66

Table D8: Model performance overview using the linear Shrinkage from Ledoit and Wolf (2022). *Return*: annualized excess return; *Total return*: the return over the entire period; *SD*: annualized standard deviation of returns; *Sortino Ratio*: downside-risk adjusted return ($\bar{r}/SD(r^-)$, where $SD(r^-)$ is the standard deviation of negative returns); *Hit Ratio*: fraction of periods with positive returns; *Sharpe Ratio*: risk adjusted returns ($r - \bar{r}_{rf}/SD(r)$); *SR CI low* and *SR CI high*: 95% confidence interval. Confidence bands are created by bootstrapping, following Ledoit and Wolf (2008): I take 1000 bootstrap samples from the return series (with replacement) and from these simulations retrieve the confidence bands of the Sharpe Ratio. For the risk-free rate r_{rf}, I use the 3-Month Treasury Bill Secondary Market Rate (Discount Basis).

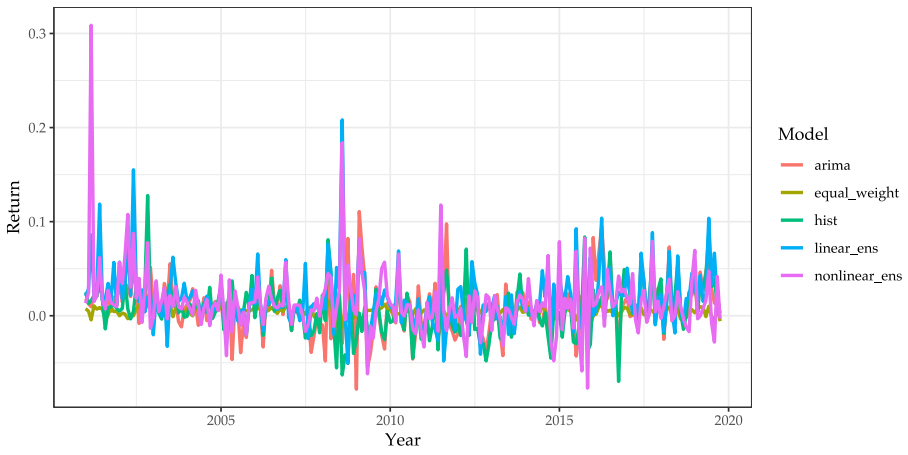


Figure D13: Monthly returns per model.

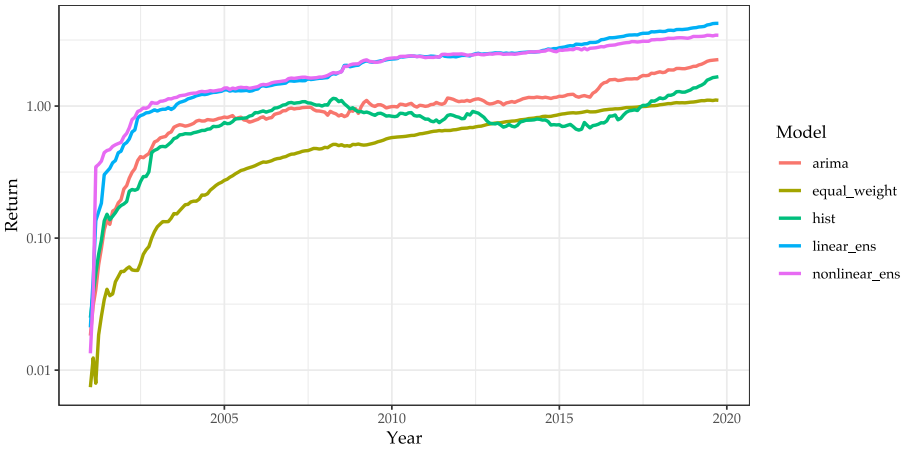


Figure D14: Cumulative monthly returns in log scale.

E Robustness checks

In this section, I include the results of different configurations that I impose to check the robustness of my results.

E.1 Transaction costs

Figure E1 shows the development of the transaction cost proxy. As expected, the half-spread increases during times of financial stress, such as the GFC. We observe that the monthly spread is mostly between one and three percent. This high value is responsible for negative returns, especially when it must be accounted for twice (when opening and closing the position).

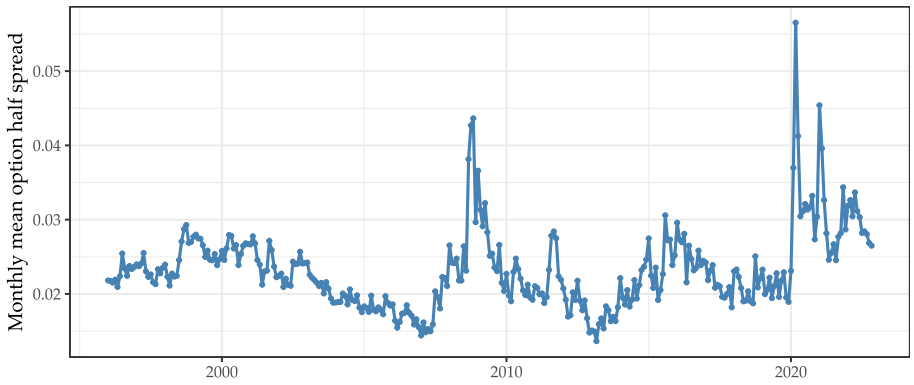


Figure E1: Monthly average option half spread. This data set includes all options in first and tenth deciles of all 28 factors. The option half spread is expressed in dollar terms.

E.2 Training periods

Using the expanding window worsens the forecasts for every model. When examining the model-suggested returns in Figure E2, two observations can be made: First, the ARMA model performs best after 2015, achieving high returns. This suggests a high autocorrelation of PC returns after 2015. Second, the performance of the ML models declines after 2015. The returns of all models except ARMA are negative until the end of 2019.

This pattern is also reflected in the models' squared prediction errors for the PCs. Figure E3 shows that the squared error increases dramatically for almost all PCs and models after 2015. This supports the hypothesis of a potential structural break prior to 2015, possibly caused by the GFC.

Table E1 shows that the prediction quality does not significantly change when excluding the GFC. For example, the linear ensemble still is able to predict factor returns better than the benchmark of zero.

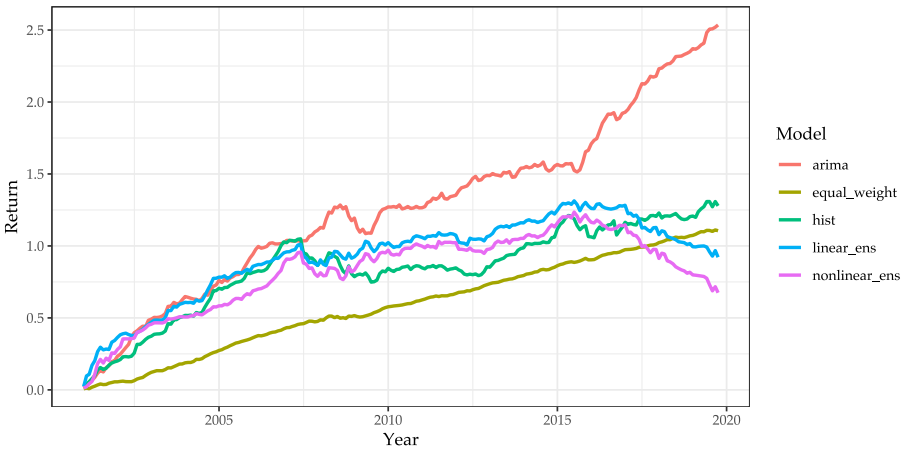


Figure E2: Cumulative monthly returns with expanding window.

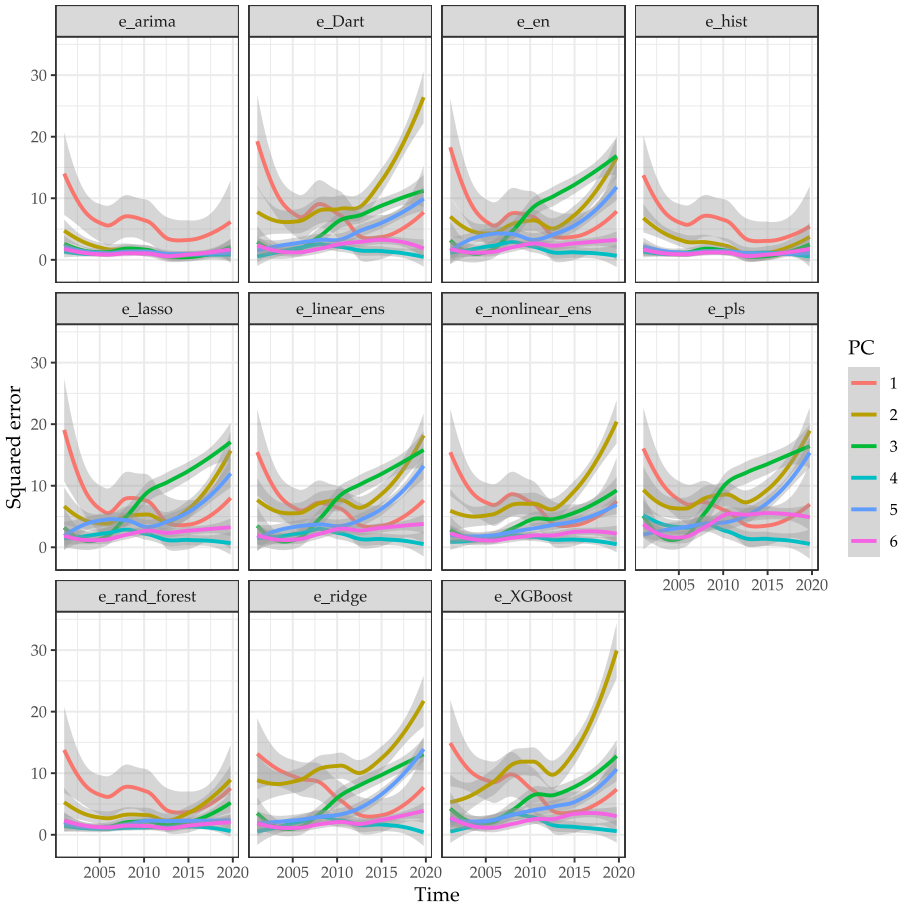


Figure E3: Squared errors per model for each PC with expanding window.

Model	PC1	PC2	PC3	PC4	PC5	PC6
arima	-0.07	0.22	-0.09	-0.06	-0.03	0.04
zero_return	0.04	0.21	-0.09	0.02	0.05	0.00
lasso	0.29	0.14	-0.04	-0.06	-0.19	-0.07
ridge	0.16	0.25	0.06	0.07	0.02	0.04
en	0.27	0.15	-0.02	-0.02	-0.10	-0.08
pls	0.07	-0.02	-0.17	-0.41	-0.10	-0.20
linear_ens	0.30	0.21	0.03	-0.00	-0.01	-0.02
rand_forest	0.16	0.18	-0.20	-0.10	-0.11	-0.17
XGBoost	0.22	0.21	-0.12	-0.11	-0.16	-0.15
Dart	0.21	0.21	-0.27	-0.04	-0.20	-0.17
nonlinear_ens	0.23	0.24	-0.12	-0.02	-0.11	-0.11

Table E1: Out-of-sample R^2 for each model, excluding the years of the GFC (2008 and 2009)

This prediction quality then also affects the overall return over the period. General return patterns do not change when excluding the GFC, as presented in Figure E4.

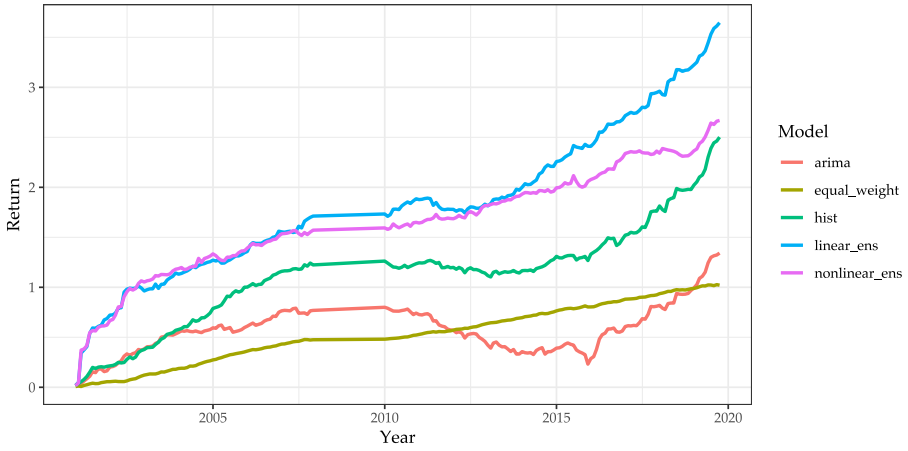


Figure E4: Squared errors per model for each PC with expanding window.

E.3 Option types

Tables E2 and E3 report the out-of-sample R^2 values for each model when using only call and only put options, respectively. Tables E4 and E5 provide the corresponding performance summaries, including annualized returns, total returns, risk-adjusted measures, and confidence intervals for the Sharpe Ratios.

Model	PC1	PC2	PC3	PC4	PC5	PC6
arima	-0.04	0.07	-0.12	0.04	-0.03	0.01
zero_return	0.03	0.17	-0.07	0.04	0.03	-0.01
lasso	0.32	0.06	-0.04	-0.27	-0.09	-0.13
ridge	0.18	0.21	0.06	-0.04	0.03	-0.00
en	0.26	0.07	-0.00	-0.39	-0.06	-0.10
pls	0.08	-0.13	-0.16	-0.47	-0.13	-0.50
linear_ens	0.32	0.16	0.03	-0.17	0.02	-0.10
rand_forest	0.21	0.10	-0.21	-0.02	-0.01	-0.20
XGBoost	0.37	0.17	-0.10	-0.06	0.00	-0.12
Dart	0.36	0.20	-0.16	-0.13	0.04	-0.15
nonlinear_ens	0.35	0.19	-0.11	-0.02	0.04	-0.12

Table E2: Out-of-sample R^2 for each model, using only call options.

Model	PC1	PC2	PC3	PC4	PC5	PC6
arima	0.02	0.00	-0.10	0.07	-0.04	-0.06
zero_return	0.04	0.15	-0.05	0.04	0.05	-0.03
lasso	0.10	-0.12	-0.11	-0.12	-0.13	-0.17
ridge	0.12	0.09	0.02	0.06	-0.02	0.01
en	0.13	-0.16	-0.09	-0.08	-0.15	-0.12
pls	-0.16	-0.21	-0.22	-0.15	-0.47	-0.21
linear_ens	0.16	0.01	0.02	0.02	-0.12	-0.07
rand_forest	0.10	-0.00	-0.04	-0.01	-0.20	-0.15
XGBoost	0.12	0.04	-0.01	0.03	-0.12	-0.06
Dart	0.14	0.05	-0.09	0.01	-0.20	-0.02
nonlinear_ens	0.16	0.06	0.00	0.05	-0.13	-0.03

Table E3: Out-of-sample R^2 for each model, using only put options.

Model	Return	Total return	SD	Sortino Ratio	Hit Ratio	Sharpe Ratio	SR CI low	SR CI high
arima	0.09	1.88	0.11	1.11	0.63	0.76	0.33	1.19
equal_weight	0.04	1.09	0.02	3.32	0.82	2.21	1.65	2.87
hist	0.05	1.14	0.11	0.65	0.61	0.44	-0.02	0.88
linear_ens	0.17	3.43	0.13	2.47	0.72	1.34	0.95	1.75
nonlinear_ens	0.18	3.57	0.13	3.35	0.69	1.34	0.99	1.75

Table E4: Model performance overview using only put options. *Return*: annualized excess return; *Total return*: the return over the entire period; *SD*: annualized standard deviation of returns; *Sortino Ratio*: downside-risk adjusted return ($\bar{r}/SD(r^-)$, where $SD(r^-)$ is the standard deviation of negative returns); *Hit Ratio*: fraction of periods with positive returns; *Sharpe Ratio*: risk adjusted returns ($r - \bar{r}_{rf}/SD(r)$); *SR CI low* and *SR CI high*: 95% confidence interval. Confidence bands are created by bootstrapping, following Ledoit and Wolf (2008): I take 1000 bootstrap samples from the return series (with replacement) and from these simulations retrieve the confidence bands of the Sharpe Ratio. For the risk-free rate r_{rf}, I use the 3-Month Treasury Bill Secondary Market Rate (Discount Basis).

Model	Return	Total return	SD	Sortino Ratio	Hit Ratio	Sharpe Ratio	SR CI low	SR CI high
arima	0.14	2.94	0.09	3.19	0.72	1.62	1.19	2.11
equal_weight	0.05	1.12	0.02	5.01	0.87	2.97	2.34	3.70
hist	0.11	2.33	0.10	1.56	0.67	1.06	0.63	1.59
linear_ens	0.21	4.19	0.10	5.13	0.78	2.01	1.62	2.53
nonlinear_ens	0.16	3.22	0.11	2.70	0.70	1.37	0.98	1.83

Table E5: Model performance overview using only put options. *Return*: annualized excess return; *Total return*: the return over the entire period; *SD*: annualized standard deviation of returns; *Sortino Ratio*: downside-risk adjusted return ($\bar{r}/SD(r^-)$, where $SD(r^-)$ is the standard deviation of negative returns); *Hit Ratio*: fraction of periods with positive returns; *Sharpe Ratio*: risk adjusted returns ($r - \bar{r}_{rf}/SD(r)$); *SR CI low* and *SR CI high*: 95% confidence interval. Confidence bands are created by bootstrapping, following Ledoit and Wolf (2008): I take 1000 bootstrap samples from the return series (with replacement) and from these simulations retrieve the confidence bands of the Sharpe Ratio. For the risk-free rate r_{rf}, I use the 3-Month Treasury Bill Secondary Market Rate (Discount Basis).

E.4 Selection of factors

Table E6 presents a performance overview of models timing the five factors with the best in-sample R^2 in the PC regression.

Model	Return	Total return	SD	Sortino Ratio	Hit Ratio	Sharpe Ratio	SR CI low	SR CI high
arima	0.31	6.07	0.15	3.49	0.74	2.02	1.55	2.59
equal_weight	0.04	1.11	0.02	3.96	0.85	2.71	2.15	3.40
hist	0.31	6.02	0.15	2.60	0.76	2.02	1.54	2.67
linear_ens	0.28	5.46	0.16	2.35	0.75	1.77	1.26	2.38
nonlinear_ens	0.35	6.89	0.14	2.85	0.83	2.51	1.92	3.30

Table E6: Model perfomance overview timing the five factors that have best in-sample R^2 in PC regression. *Return*: annualized excess return; *Total return*: the return over the entire period; *SD*: annualized standard deviation of returns; *Sortino Ratio*: downside-risk adjusted return ($\bar{r}/SD(r^-)$, where $SD(r^-)$ is the standard deviation of negative returns); *Hit Ratio*: fraction of periods with positive returns; *Sharpe Ratio*: risk adjusted returns ($r - \bar{r}_{rf}/SD(r)$); *SR CI low* and *SR CI high*: 95% confidence interval. Confidence bands are created by bootstrapping, following Ledoit and Wolf (2008): I take 1000 bootstrap samples from the return series (with replacement) and from these simulations retrieve the confidence bands of the Sharpe Ratio. For the risk-free rate r_{rf}, I use the 3-Month Treasury Bill Secondary Market Rate (Discount Basis).

E.5 Characteristics construction

Table E7 and E8 show the out-of-sample R^2 when leaving out the approach of Kagkadis et al. (2024) and leaving out factor characteristics altogether, respectively.

The return summaries when leaving out dimensionality reduction of Kagkadis et al. (2024) and without characteristics are represented in Table E9 and E10, respectively.

Model	PC1	PC2	PC3	PC4	PC5	PC6
arima	0.03	0.17	-0.09	-0.00	-0.02	0.03
zero_return	0.04	0.17	-0.08	0.02	0.01	-0.02
lasso	0.22	0.07	-0.01	0.03	-0.12	-0.22
ridge	0.17	0.17	0.07	0.06	0.05	-0.05
en	0.15	0.06	-0.03	-0.02	-0.09	-0.11
pls	0.06	-0.02	-0.19	-0.25	-0.20	-0.47
linear_ens	0.26	0.15	0.05	0.05	-0.02	-0.13
rand_forest	0.19	0.14	-0.03	-0.15	-0.11	-0.18
XGBoost	0.23	0.23	0.01	-0.07	-0.13	-0.19
Dart	0.27	0.16	-0.00	-0.13	-0.13	-0.14
nonlinear_ens	0.27	0.22	0.06	-0.06	-0.08	-0.11

Table E7: Out-of-sample R^2 for each model, leaving out dimensionality reduction following Kagkadis et al. (2024).

Model	PC1	PC2	PC3	PC4	PC5	PC6
arima	0.03	0.17	-0.09	-0.00	-0.02	0.03
zero_return	0.04	0.17	-0.08	0.02	0.01	-0.02
lasso	0.27	0.04	-0.09	-0.06	-0.16	-0.16
ridge	0.10	0.11	-0.05	-0.00	-0.01	-0.10
en	0.28	0.03	-0.14	-0.06	-0.09	-0.11
pls	0.11	-0.09	-0.24	-0.28	-0.17	-0.47
linear_ens	0.26	0.09	-0.04	-0.02	-0.05	-0.14
rand_forest	0.18	0.11	-0.04	-0.15	-0.12	-0.18
XGBoost	0.30	0.20	-0.07	-0.13	-0.15	-0.16
Dart	0.31	0.12	-0.08	-0.20	-0.14	-0.16
nonlinear_ens	0.31	0.20	0.02	-0.10	-0.08	-0.10

Table E8: Out-of-sample R^2 for each model, excluding all characteristics in the explanatory variables.

Model	Return	Total return	SD	Sortino Ratio	Hit Ratio	Sharpe Ratio	SR CI low	SR CI high
arima	0.11	2.25	0.10	2.03	0.67	1.10	0.67	1.58
equal_weight	0.04	1.11	0.02	3.96	0.85	2.71	2.11	3.42
hist	0.07	1.67	0.09	1.44	0.66	0.81	0.38	1.24
linear_ens	0.19	3.79	0.11	4.35	0.76	1.79	1.43	2.21
nonlinear_ens	0.17	3.56	0.12	3.50	0.73	1.43	1.08	1.88

Table E9: Model perfomance overview leaving out the dimensionality reduction in characteristics weights, following Kagkadis et al. (2024). *Return*: annualized excess return; *Total return*: the return over the entire period; *SD*: annualized standard deviation of returns; *Sortino Ratio*: downside-risk adjusted return ($\bar{r}/SD(r^-)$, where $SD(r^-)$ is the standard deviation of negative returns); *Hit Ratio*: fraction of periods with positive returns; *Sharpe Ratio*: risk adjusted returns ($r - \bar{r}_{rf}/SD(r)$); *SR CI low* and *SR CI high*: 95% confidence interval. Confidence bands are created by bootstrapping, following Ledoit and Wolf (2008): I take 1000 bootstrap samples from the return series (with replacement) and from these simulations retrieve the confidence bands of the Sharpe Ratio. For the risk-free rate r_{rf}, I use the 3-Month Treasury Bill Secondary Market Rate (Discount Basis).

Model	Return	Total return	SD	Sortino Ratio	Hit Ratio	Sharpe Ratio	SR CI low	SR CI high
arima	0.11	2.25	0.10	2.03	0.67	1.10	0.65	1.53
equal_weight	0.04	1.11	0.02	3.96	0.85	2.71	2.11	3.36
hist	0.07	1.67	0.09	1.44	0.66	0.81	0.37	1.26
linear_ens	0.23	4.55	0.13	4.41	0.81	1.81	1.49	2.33
nonlinear_ens	0.18	3.59	0.11	3.20	0.77	1.67	1.29	2.04

Table E10: Model perfomance overview excluding the characteristics weights from the predictions. *Return*: annualized excess return; *Total return*: the return over the entire period; *SD*: annualized standard deviation of returns; *Sortino Ratio*: downside-risk adjusted return ($\bar{r}/SD(r^-)$, where $SD(r^-)$ is the standard deviation of negative returns); *Hit Ratio*: fraction of periods with positive returns; *Sharpe Ratio*: risk adjusted returns ($r - \bar{r}_{rf}/SD(r)$); *SR CI low* and *SR CI high*: 95% confidence interval. Confidence bands are created by bootstrapping, following Ledoit and Wolf (2008): I take 1000 bootstrap samples from the return series (with replacement) and from these simulations retrieve the confidence bands of the Sharpe Ratio. For the risk-free rate r_{rf}, I use the 3-Month Treasury Bill Secondary Market Rate (Discount Basis).

F AI usage disclosure

In this section, I describe how I used artificial intelligence tools in this master's thesis.

Which AI tools were used and how? I used ChatGPT and Claude. Both served two purposes: (1) they supported me as advisors in developing and refining my R code, and (2) they helped me improve the spelling, grammar, and wording of the written text. When working with R, I did not copy any code generated by those tools. Since the suggestions were sometimes inconsistent, I mainly used the tool to test whether my own code worked as intended and to get inspiration for performance improvements. For the written thesis, I used AI to correct selected passages, check for comprehensibility, and restructure overly complicated sentences. Since AI occasionally neglects important information when rephrasing texts, I did not insert its results directly, but used its suggestions to improve readability. Since the thesis was written in LaTeX, I also used ChatGPT and Claude to assist with formatting questions.

To what extent did these tools contribute to improving the quality of the thesis? The targeted use of proofreading, limited rephrasing suggestions and coding aids helped me to improve the clarity and readability of the text while ensuring that the R code I wrote worked as expected. These applications contributed to a more polished final document without affecting the content of the thesis.

What potential risks were identified in the use of AI, and what measures were taken to mitigate these risks? Since I only relied on AI for linguistic revision, coding guidance, and formatting assistance, there were no risks in terms of content. Nevertheless, AI tools can make mistakes or overlook information. To avoid this, I double-checked every suggestion – whether linguistic, structural or coding-related – before adopting it.

What insights were gained from using AI tools when writing the thesis? As a non-native speaker dealing with complex academic material, I found AI-based proofreading particularly helpful in ensuring correct and clear English. This experience also demonstrated the importance of using AI critically: it can provide valuable inspiration and support, but must be supplemented by careful human judgement at every step.