Variable selection for minimum-variance portfolios[*]

Guilherme V. Moura André P. Santos Hudson S. Torrent

November 28, 2024

Abstract

Machine learning (ML) methods have successfully identified predictor variables for the equity premium of individual stocks. In this paper, we investigate whether ML can also select predictors for optimal portfolio choice. We parameterize portfolio weights as a function of a large pool of firm characteristics, including non-linear transformations, yielding 4,610 predictors. Employing ML to select predictors significantly enhances portfolio performance: minimum-variance portfolios achieve lower risk compared to sparse specifications, especially when non-linear terms are included. Moreover, some ML methods select predictors that reduce portfolio risk and enhance returns, resulting in minimum-variance portfolios with strong performance in terms of Sharpe ratios.

Keywords: Ensembles, factor zoo, regularization, horseshoe prior.

JEL classification: B26; C58; G11.

[*] Guilherme V. Moura is with Universidade Federal de Santa Catarina, e-mail: `guilherme.moura@ufsc.br`. André A. P. Santos is with CUNEF Universidad, e-mail: `andre.santos@cunef.edu` (corresponding author). Hudson S. Torrent is with Universidade Federal do Rio Grande do Sul, e-mail: `hudson.torrent@ufrgs.br`. We thank Christos Konstantinous, Alexandre Rubesam, Rainer Schüssler, Rasmus Lonn, Robinson Reyes, and participants of the 5th International Workshop in Financial Econometrics 2023, the Financial Econometrics meets Machine Learning (FinEML) Conference 2023, the Economics of Financial Technology Conference 2024, and the Finance Forum 2024 for useful comments. A previous version of this paper circulated under the title "High-dimensional variable selection for portfolio optimization: Is machine learning helpful?". A.P. Santos gratefully acknowledge the support of the Agencia Estatal de Investigación (Spain) under grant PID2022-138289NB-I00 and Comunidad Autonoma de Madrid Government through project 2022-T1/SOC-24167 (Convocatoria Talento).

1 Introduction

Which variables are important to the portfolio construction problem, and how can they be selected? Addressing this question is difficult considering the availability of a myriad of predictors that seem to explain the cross-sectional patterns of stock returns (see, for instance, Cochrane, 2011; Harvey, Liu, and Zhu, 2016; McLean and Pontiff, 2016; Hou, Xue, and Zhang, 2020). The problem becomes even more challenging when non-linear transformations, such as interactions between covariates and high-order moments, are taken into account, which results in an exponential increase in the predictor space. In this context, the number of features can easily surpass the number of data points, and the application of standard statistical methods, such as OLS, is known to perform poorly, see Nagel (2021). In this paper, we assemble a data set with 4,610 firm-level predictors and demonstrate that machine learning (ML) variable selection methods are capable of selecting subsets of these features that are important for the portfolio construction problem.

Much of the literature on variable selection in empirical asset pricing considers ad-hoc sparse specifications based on a few predetermined state variables in order to explain the cross-section patterns of stock returns (e.g. Fama and French, 1993, 2015; Hou, Xue, and Zhang, 2015) and to determine how they affect optimal portfolio allocations (Aït-Sahalia and Brandt, 2001). Kozak, Nagel, and Santosh (2020) refer to this approach as "characteristics-sparse models". For instance, Chan, Karceski, and Lakonishok (1999) and Brandt, Santa-Clara, and Valkanov (2009) propose portfolio allocation methods based on low-dimensional factor specifications containing three characteristics. An obvious limitation of this approach is that many important variables can be omitted. Two notable and recent exceptions are the works of DeMiguel, Martin-Utrera, Nogales, and Uppal (2020) and Kozak et al. (2020) who employ high-dimensional data sets containing a large number of characteristics and find that the number of relevant ones for the portfolio construction problem is larger in comparison to the standard specifications based on only a few ad-hoc characteristics. Nagel (2021) argue that "One interpretation of this expansion in the number of factors is that the literature is slowly adjusting to the fact that there are, indeed, relevant omitted factors."

In this paper, we present a methodological framework for high-dimensional predictor selection and provide empirical evidence that enables us to establish four main conclusions that are useful for academics and market practitioners. First and foremost, the choice of variable selection method matters: there are substantial differences in terms of the *preference* of each method with respect to specific features and to the *number* of features each method selects. Second, the number of selected features that are relevant for the portfolio construction problem exceeds the number of variables used in popular low-dimensional factor models. Third, augmenting the predictor space with non-linear transformations of the original characteristics leads to optimal portfolios with improved performance.[1] Fourth, working with a subset of important predictor variables results in minimum-variance portfolios that outperform their benchmarks in terms of a lower standard deviation of out-of-sample portfolio returns.

We add to the existing literature by extending the works of Kozak et al. (2020) and DeMiguel et al. (2020) along two dimensions. First, we raise the statistical challenge by assembling a set of predictor variables containing the 95 firm characteristics considered in Gu et al. (2020) and augment it with the second-order and cross-product (interaction) terms among all characteristics, yielding a total of 4,610 predictors. This allows us to study not only *whether* and to *what extent* non-linear functions of features that are important for portfolio construction, but also *what type* non-linearity (i.e. non-linearities of individual predictors or interactions between covariates) that are most important in helping the investor improve portfolio performance. Existing empirical evidence (Moritz and Zimmermann, 2016; Bryzgalova, Pelger, and Zhu, 2020; Chen, Pelger, and Zhu, 2020; Freyberger, Neuhierl, and Weber, 2020; Gu, Kelly, and Xiu, 2020, 2021) suggests that non-linear functions of firm characteristics can be an important source *of predictability of security returns.* Thus, it is natural to argue whether non-linearities are relevant and what type they may be *from a portfolio choice perspective.*

Our second contribution is to expand the set of variable selection methods used to identify the most important predictors. Specifically, we consider three classes of variable selection methods: *regularization*, *ensembles*, and *Bayesian* methods. While regularization methods perform variable selection by shrinking the coefficients using the $L1$ regularization (or a

[1] Non-linear transformations have been extensively studied in the asset pricing literature. For instance, Harvey and Siddique (2000) assume that the stochastic discount factor is quadratic in the market return. Hong, Lim, and Stein (2000) show that momentum interact with firm size and analyst coverage: momentum strategies declines with firm size and work better among stocks with low analyst coverage. Medhat and Schmeling (2022) show that previous month's return and share turnover also interact: low-turnover stocks exhibit significant short-term reversal whereas high-turnover stocks exhibit short-term momentum.

combination of $L1$ and $L2$ regularizations), the ensemble method considered in the paper operates via a stagewise additive modeling procedure that iteratively estimates the model by sequentially adding new components, with early stopping of the algorithm to avoid overfitting. The Bayesian variable selection, on the other hand, performs regularization by specifying a prior over the parameters, which shrinks weak signals toward zero while allowing strong signals to remain unshrunk. Finally, as a robustness check, we implement the marginal screening method of Fan and Lv (2008), which is computationally efficient and suitable for ultra-high dimensions of predictors.

It is also worth highlighting three important methodological aspects of our work. First, consistent with our interest in the portfolio choice problem, we focus on *directly* parameterizing the objective of interest - *portfolio weights* - as a function of lagged firm-level characteristics and their cross-products and second-order transformations, instead of predicting *stock returns.* This difference is important because variables that lead to better predictions of individual security returns in terms of higher out-of-sample R-squared do not necessarily lead to better portfolios in terms of standard metrics such as the out-of-sample standard deviation of portfolio returns or the portfolio's Sharpe ratio; see Nagel (2021) for a discussion. The reason is that portfolio performance depends on the properties of the covariance matrix of asset returns, and R-squared measures are uninformative about those properties.[2]

A second important aspect of our methodology is the use of the parametric portfolio policy approach of Brandt et al. (2009) to parameterize portfolio weights as a linear function of non-linear transformations of lagged firm characteristics. We borrow analytical results obtained in DeMiguel et al. (2020) and reformulate the parametric portfolio approach as a penalized regression model that allows us to study the role played by alternative variable selection strategies in identifying relevant predictors for the portfolio construction problem.

Finally, the third important aspect of our methodology is the focus on minimum-variance portfolios. Several reasons motivate this choice. First, minimum-variance strategies are among

[2] Pástor and Stambaugh (2000) show that a model that is better for pricing is not necessarily better for investing, because investors are usually subject to margin requirements and model uncertainty that prevent them from implementing certain investment strategies suggested by asset pricing models. Fabozzi, Huang, Jiang, and Wang (2024) builds on the approach developed by Pástor and Stambaugh (2000) and find that, although the 4-factor model of Hou, Xue, and Zhang (2015) is better than the 5-factor model of Fama and French (2015) at explaining anomalies, the two models have similar performance in terms of investing.

the most popular smart-beta strategies in the U.S. (Ang, Madhavan, and Sobczyk, 2017).[3] Second, minimum-variance portfolios are less sensitive to estimation errors in expected returns and often lead to robust performance in practice (DeMiguel and Nogales, 2009). Third, the minimum-variance parametric portfolios can be cast as an unconstrained regression problem, which allows for the deployment of a wide range of variable selection methods in a natural way. Finally, arbitrary characteristic sparsity can be particularly harmful to these portfolios. This is the case because a potentially high number of predictors can help reduce portfolio risk, even if they do not contribute to increasing portfolio mean returns. For instance, a predictor with a low covariance with the benchmark portfolio or with other predictors is likely to be important for constructing a minimum-variance portfolio, even if it does not aid in increasing average portfolio returns. We circumvent the arbitrary sparsity by considering a very large cross-section of firm characteristics and employing variable selection methods to identify predictors that are relevant for the minimum-variance construction problem.

We perform an empirical exercise in which we use the selected predictors from each variable selection method to construct minimum-variance portfolios for the universe of NYSE, NASDAQ, and NYMEX stocks and conduct an out-of-sample evaluation. The results point to a clear superiority of the portfolios obtained with the variable selection methods in comparison to those obtained with ad hoc sparse specifications such as the 3- and 5-factor models of Fama and French (1993, 2015) as well as classic benchmark strategies such as the value-weighted and equally weighted portfolios. For instance, the minimum-variance portfolios with predictors selected using the best-performing method delivered out-of-sample returns with an annualized standard deviation of 9.1%, and this result is further improved to 8.3% when using the augmented data set with all non-linear predictor transformations. These figures are substantially and statistically lower than those obtained with the 3-factor (15%) and 5-factor (14.6%) models.

We also investigate the importance of individual predictors. We find that the main effects and interactions among market beta, different measures of return volatility, return momentum, liquidity, and bid-ask spread are among the most important predictors. We provide an interpretation of the predictor importance by i) identifying which aspects of the investor's utility the predictors contribute to the most, and ii) showing how predictor importance translates to

[3] A total of 18 minimum-volatility ETFs are traded in the U.S. They have $52.08 billion in assets under management as of July 2024 (see `etf.com/topics/low-volatility`).

portfolio allocations. For instance, stocks with lower liquidity receive higher weights, and this result is stronger for stocks with lower idiosyncratic volatility. Stocks with lower market betas receive higher weights, and this result is stronger for stocks with a higher standard deviation of liquidity. Our results are in line with those reported in Scherer (2011), which finds that risk-based characteristics such as market beta and idiosyncratic volatility can explain the allocations of minimum-variance portfolios. Our paper extends those results and shows that different measures of size, return momentum, and liquidity are also important, as well as their interactions with different risk-based characteristics.

To assess the economic significance of the most important predictors, we examine whether variations in their estimated parametric-portfolio coefficients are associated with fundamental state variables that reflect market conditions, such as investor sentiment, market-wide liquidity, and economic recession. Specifically, we use the investor sentiment index from Baker and Wurgler (2006), the liquidity index from Pástor and Stambaugh (2003), and the NBER recession indicator. Our analysis shows that variations in the estimated coefficients of the most influential predictors are indeed related to these state variables. Furthermore, the variations in coefficients associated with the main effects of predictors and the interactions between predictors tend to be linked to different state variables, which suggests that different classes of predictors capture distinct market attributes.

Our paper is connected to a growing literature that employs machine learning methods to support portfolio choice decisions; see Kelly and Xiu (2023) for a recent review of the literature. Li and Rossi (2020), DeMiguel, Gil-Bazo, Nogales, and Santos (2023), and Kaniel, Lin, Pelger, and Van Nieuwerburgh (2023) employ ML methods to construct prediction-based portfolios of mutual funds, whereas Gu et al. (2020) focuses on prediction-based stock portfolios. ML methods have also been used in economic and financial forecasting. For instance, Rapach, Strauss, and Zhou (2013), Chinco, Clark-Joseph, and Ye (2019), and Freyberger, Neuhierl, and Weber (2020) use the lasso and the elastic net to select regressors for stock return prediction, whereas Dong, Li, Rapach, and Zhou (2022) employs ML to predict the aggregate market excess return. Bai and Ng (2009), Buchen and Wohlrabe (2011), Döpke, Fritsche, and Pierdzioch (2017), and Kauppi and Virtanen (2021) use the boosting method for macroeconomic forecasts. A common denominator across these studies is that employing variable selection methods leads to improved results compared to those obtained with standard statistical methods.

Our paper is mostly connected with the works of Simon, Weibels, and Zimmermann (2022) and Lamoureux and Zhang (2024). Simon et al. (2022) formulates a mean-variance parametric portfolio policy using a non-linear functional form based on deep neural networks. In our paper, we preserve the linear functional form and model non-linearities via second-order and cross-product variable transformations. While the approach of Simon et al. (2022) can capture non-linearities in a flexible way, our method is highly interpretable and allows us to understand not only which *how many* predictors survive the different variable selection processes but also how *why* a selected predictor is important according to a given method. Lamoureux and Zhang (2024) uses the parametric portfolio approach to study whether an investor with power utility can exploit the predictability contained in a set of six characteristics. Our paper, on the other hand, focuses on minimum-variance portfolios and considers a much larger cross-section of predictors, which allows us to understand the extent to which ad-hoc sparsity is detrimental to portfolio construction.

The rest of the paper is organized as follows. Section 2 describes the dataset used in the paper. Section 3 details the portfolio strategies as well as the methods for variable selection and regularization. Section 4 presents the results of an empirical application. Finally, Section 5 concludes.

2 Data

We use the dataset of 95 monthly firm characteristics utilized in Gu et al. (2020). The database was downloaded from Dacheng Xiu's homepage (`https://dachxiu.chicagobooth.edu/download/datashare.zip`).[4] Stock returns are sourced from CRSP. Our sample spans from January 1970 to June 2019 (a total of 594 months).

We perform feature engineering to augment the original dataset containing the main effects of the 95 predictors. Specifically, for each firm characteristic, we obtain the square values and

[4] Table A.6 of the supplementary material of Gu et al. (2020) brings additional details of the characteristics. The supplementary material can be download at `https://dachxiu.chicagobooth.edu/download/ML_supp.pdf`. Although the original data set used in Gu et al. (2020) contains 94 characteristics, the data set available Dacheng Xiu's homepage contains one additional characteristic, market value of equity ($mve0$). Although $mve0$ is highly correlated with another measure of firm size in the data set ($mvel1$), we opted to retain both variables.

cross-products among all predictors; that is, we define $pred^2_{i,t}$ and $pred_{i,t} \times pred_{j,t}$ for $i \neq j$ to denote the second-order and interaction effects, respectively.[5] The augmented dataset contains the predictors' main effects, the second-order effects, and the interaction effects. The final set contains 4,610 predictors. The total number of firms in our sample is 15,701, with the fewest firms in January 1973 (1,443 firms) and the most firms in December 1997 (5,128 firms).

To alleviate the impact of extreme observations, we cross-sectionally winsorize each predictor at the 1st and 99th cross-sectional percentiles, as in Green, Hand, and Zhang (2017). We also follow Brandt et al. (2009) and standardize each predictor so that its cross-sectional mean is zero and its standard deviation is one. Missing characteristic values are set to zero. The resultant standardized predictor is a long-short portfolio that goes long on stocks whose characteristics are above the cross-sectional average and short on stocks whose characteristics are below the cross-sectional average.

We report in Table 1 descriptive statistics of the predictor returns, which are defined as

$$rpred_{t+1} = \frac{pred_t^{\intercal} r_{t+1}}{N_t},$$

where $pred_t$ is the standardized predictor long-short portfolio at time t, r_{t+1} is the vector of stock returns at time $t+1$, and N_t is the number of stocks at time t. The Table reports the average monthly return, the standard deviation of monthly returns, the 25th and 75th percentiles, and the correlation of the predictor return with the equally weighted and value-weighted portfolio returns. The first four rows report aggregate values across i) all predictors, ii) predictors' main effects, iii) predictor's second-order effects, and iv) predictors' interaction effects. The remaining rows report individual statistics for the top and bottom 10 predictors within each category, sorted in terms of average monthly return.

The overall average monthly return of all predictors is slightly negative at -0.003%, with a standard deviation of 0.523% and a near-zero correlation with the equally weighted and value-weighted portfolio returns. We find similar results for the subgroups of main effects, second-order effects, and interaction effects, except that second-order effects display higher

[5] The predictors Convertible debt indicator (*convind*), Dividend initiation (*divi*), Dividend omission (*divo*), R&D increase (*rd*), Secured debt indicator (*securedind*), and Sin stocks (*sin*) are defined as binary variables and therefore we do not consider their second-order effects. Additionally, the squared value of the predictor *beta* is not computed since its second-order effect is represented by the predictor *betasq*.

correlations with the equally weighted and value-weighted portfolio returns. The analysis of the top and bottom 10 predictors within each category, on the other hand, reveals a wide variation in performance. For instance, among the top-10 main effects, industry momentum (*indmom*) shows the highest average return of 0.3%, while book-to-market (*bm*) shows the lowest at 0.1%. The interaction between illiquidity and size ($ill \times mve0$) earns the highest average return (0.9%), whereas the interaction between short-term reversal and financial statement score ($mom1m \times ps$) earns the lowest (-0.4%). We also observe substantial variation (between -0.4 and 0.6) in the correlations of the top and bottom 10 predictors with the equally weighted and value-weighted portfolio returns.

3 Methods

We now describe the portfolio optimization framework considered in the paper, as well as the ML methods used to identify predictors that are helpful for the portfolio minimum-variance construction problem. First, we review the parametric portfolio policy of Brandt et al. (2009) and its minimum-variance reformulation as an unconstrained regression problem. Second, we discuss how alternative variable selection methods can be introduced into this framework. Finally, we discuss the approach to interpret the importance of the selected predictors.

The parametric policy is a portfolio optimization task aimed at maximizing utility. This is achieved by adjusting a benchmark portfolio (such as equally weighted or value-weighted portfolios) using a series of asset-specific predictors to enhance the investor's utility. Specifically, the benchmark portfolio is modified by adding a weighted sum of long-short portfolios. These portfolios are derived from K predictors that are normalized cross-sectionally to have a mean of zero and a standard deviation of one. The parametric policy strictly involves equity investments, explicitly excluding any allocation to risk-free assets.

The composition of the parametric portfolio at a given time t, denoted as $w_t(\theta) \in \mathbb{R}^{N_t}$, is expressed as follows:

$$w_t(\theta) = w_{b_t} + \frac{1}{N_t}\sum_{k=1}^{K} x_{k,t}\theta_k = w_{b_t} + \frac{X_t\theta}{N_t}, \tag{1}$$

where $\theta^\intercal = (\theta_1, \theta_2, \ldots, \theta_K) \in \mathbb{R}^K$ is a $1 \times K$ vector of coefficients, $X_t = [x_{1,t}, x_{2,t}, \ldots, x_{K,t}] \in \mathbb{R}^{N_t \times K}$ represents a matrix of asset characteristics at time t, $w_{b,t} \in \mathbb{R}^{N_t}$ is the benchmark portfolio at time t, $x_{k,t} \in \mathbb{R}^{N_t}$ is the standardized long-short portfolio for the kth predictor, θ_k is the associated coefficient for the kth predictor, and N_t is the total number of stocks at time t. The standardization ensures that the average of $x_{k,t}\theta_k$ across all assets is zero, leading to an optimal portfolio where deviations from the benchmark weights sum to zero, and the total portfolio weights are always equal to one.

The return of this parametric portfolio at a subsequent time point, $t + 1$, denoted as $r_{p,t+1}$, is given by

$$r_{p,t+1}(\theta) = r_{b,t+1} + \theta^\intercal r_{c,t+1}, \tag{2}$$

where $r_{t+1} \in \mathbb{R}^{N_t}$ is the return vector at time $t + 1$, $r_{b,t+1} = w_{b,t}^\intercal r_{t+1}$ represents the benchmark portfolio's return, and $r_{c,t+1} = \frac{X_t^\intercal r_{t+1}}{N_t}$ is the predictor return vector at that time. This predictor return vector encapsulates the returns from the long-short portfolios related to the K predictors, adjusted by the number of assets, N_t. The equation reveals that the return on the parametric portfolio is a sum of the benchmark return and the yield from a linear combination of the characteristic portfolios.

3.1 Minimum-variance parametric portfolios

A very large body of literature on portfolio optimization considers the minimum-variance policy; see, for instance, Green and Hollifield (1992), Jagannathan and Ma (2003), Clarke, De Silva, and Thorley (2006), DeMiguel and Nogales (2009), and Clarke, De Silva, and Thorley (2011) just to name a few. This strategy performs well as it is more robust than the mean-variance policy since it does not require estimating mean returns, which is a notoriously difficult task.

We assume a minimum-variance investor who solves the following problem:

$$\min_{\theta} \ \frac{1}{2}\text{var}\left[r_{p,t+1}(\theta)\right] \equiv \frac{1}{2}E\left[(\dot{r}_{p,t+1}(\theta))^2\right], \tag{3}$$

where $\dot{r}_{p,t+1}(\theta)$ is the portfolio return vector centered on zero, i.e. $\dot{r}_{p,t+1}(\theta) = r_{p,t+1}(\theta) - \bar{r}_p(\theta)$ and $\bar{r}_p(\theta)$ is the mean portfolio return. In practice, one minimizes the empirical version of eq. (3) using a sample with T observations, defined as

$$\min_{\theta} \ \frac{1}{2}\frac{1}{(T-1)}\sum_{t=1}^{T-1}\left(\dot{r}_{b,t+1}+\theta^{\intercal}\dot{r}_{c,t+1}\right)^{2}, \tag{4}$$

where $\dot{r}_{b,t+1}$ and $\dot{r}_{c,t+1}$ are, respectively, the mean-centered benchmark return and predictor return. It is straightforward to note that the portfolio loss function in (4) can be formulated as a regression problem. Specifically, the optimal parameter θ^{*} in eq. (4) can be estimated with a time-series regression without an intercept,

$$\dot{r}_{b,t} = -\theta^{\intercal}\dot{r}_{c,t} + \epsilon_{t}, \tag{5}$$

where ϵ_t is the error term. The corresponding ordinary least squares (OLS) solution of (4) is given by

$$\hat{\theta} = -\hat{\Sigma}_{c}^{-1}\hat{\sigma}_{bc}, \tag{6}$$

where $\hat{\Sigma}_c$ is the sample covariance matrix of the predictor-return vector r_c, and $\hat{\sigma}_{bc}$ is the sample vector of covariances between the benchmark portfolio return r_b and the predictor-return vector r_c.[6]

The OLS solution of the empirical minimum-variance portfolio problem in (6) suffers from a major drawback: when $K > T$, that is, when the number of predictors exceeds the number of data points, the OLS solution is poor or unfeasible since the sample covariance matrix $\hat{\Sigma}_c$ is no longer positive definite. This poses a serious limitation to the use of standard regression methods in estimating the parameters of the minimum-variance parametric portfolios when the dimension of the predictor space is high.

[6] The regression formulation of the parametric minimum-variance portfolio problem in (5) differs from existing portfolio-regression formulations. For instance, Britten-Jones (1999) shows that the tangency portfolio can be obtained by a regression of a constant $\mathbf{1}$ onto a set of asset's excess returns, without an intercept term. Kempf and Memmel (2006) show that the global minimum-variance portfolio can be obtained by regressing the returns of a given asset i onto the differences between the returns of asset i and all other assets. The portfolio-regression formulation adopted in the paper is closer to that developed in DeMiguel et al. (2020) as it is based on the parametric-portfolio formulation.

3.2 Variable selection methods

We will now discuss the three classes of variable selection methods that help solve the empirical minimum-variance parametric portfolio problem in eq. (4) in situations where the number of predictors can exceed the number of data points. The three major classes of variable selection procedures considered are regularization methods, Bayesian methods, and ensemble methods.

3.2.1 Regularization

We consider a formulation of the unconstrained regression in (4) in which a regularization term is added to encourage sparsity,

$$\min_{\theta} \ \frac{1}{2}\frac{1}{(T-1)}\sum_{t=1}^{T-1}\left(\dot{r}_{b,t+1}+\theta^{\intercal}\dot{r}_{c,t+1}\right)^2+\Omega_{\lambda}\left(\theta\right), \tag{7}$$

where $\Omega_{\lambda}\left(\theta\right)$ defines the penalization function and λ denotes the penalization hyperparameters. Next, we describe the five penalty-based methods used in the paper.

Ridge Hoerl and Kennard (1970) proposed the ridge estimator to reduce mean squared error of the OLS at the cost of some bias. The idea is to add a penalty proportional to the *squared magnitude* of the coefficients. The ridge penalization is defined as

$$\Omega_{\lambda}\left(\theta\right)=\lambda\sum_{k=1}^{K}\theta_k^2, \tag{8}$$

where λ is the regularization hyperparameter that controls the strength of the penalty. The decrease in mean squared error compared to the OLS occurs for some intermediate value of λ, for which the reduction in variance of $\widehat{\theta}$ that solves (8) surpasses the bias induced by the regularization. Ridge regression offers the advantage of a computationally straightforward analytic solution. However, in this approach, coefficients associated with less relevant predictors are gradually shrunk toward zero but never precisely reach it. Consequently, ridge regression is not well suited for predictor selection (see, for example, Masini, Medeiros, and Mendes, 2023).

Least absolute sum of squares operator (lasso) With the objective of performing variable selection jointly with regularization, Tibshirani (1996) proposed the lasso, where the penalty is the *absolute magnitude* of the coefficients. The formulation becomes

$$\Omega_\rho\left(\theta\right) = \rho \sum_{k=1}^{K} |\theta_k|, \tag{9}$$

where ρ is the regularization hyperparameter. Due to the specific nature of the constraint in the lasso, reducing ρ sufficiently will result in some coefficients being precisely zero. Consequently, the lasso exhibits a form of continuous subset selection, providing a *sparse* estimator of the paremeter vector θ (see, for example, Hastie, Tibshirani, Friedman, and Friedman, 2009).

Adaptive lasso (adalasso) Although the lasso is a consistent method for variable selection under certain conditions, there exist scenarios where the lasso is inconsistent for this purpose. Zou (2006) proposed the adalasso regression, which is a modified version of the lasso regression based on adaptive weights used to penalize coefficients differently,

$$\Omega_{\rho,\phi}\left(\theta\right) = \rho \sum_{k=1}^{K} \phi_k |\theta_k|, \tag{10}$$

where ϕ_k are weights typically set as the inverse of the absolute values of the estimates from an initial ridge regression, and ρ is the regularization parameter. As the lasso, the adalasso also delivers sparsity and efficient estimation algorithm, but enjoys the oracle property (Fan and Li, 2001), meaning that it has the same asymptotic distribution as the OLS conditional on knowing the variables that should enter the model.

Elastic net Introduced by Zou and Hastie (2005), the elastic net strategically combines the merits of both lasso and ridge regression, and the penalization function becomes

$$\Omega_{\lambda,\rho}\left(\theta\right) = \lambda\rho\ \sum_{k=1}^{K} |\theta_k| + \lambda(1-\rho)\ \sum_{k=1}^{K} \theta_k^2. \tag{11}$$

The $L1$-norm term ($\lambda\rho\ \sum_{k=1}^{K} |\theta_k|$) can be used to control the sparsity of the estimated parameter vector θ and the $L2$-norm term ($\lambda(1-\rho)\ \sum_{k=1}^{K} \theta_k^2$) to increase its stability. For the case with $\rho = 0$, the objective function includes only the $L2$-norm term, and thus, elastic net is equivalent to ridge regression. If, on the other hand, $\rho = 1$, the objective function includes only the $L1$-norm term, and lasso regression is performed. The elastic net regression can offer advantages over using either lasso or ridge alone, particularly in the presence of correlated predictors, when it outperforms the lasso (see, for example, Zou and Zhang, 2009).

Smoothly clipped absolute deviation (SCAD) The SCAD penalization introduced by Fan and Li (2001) combines the strengths of both ridge and lasso by offering a penalty that varies with the coefficient's magnitude. The objective function can be expressed as

$$\Omega_p(\theta) = \sum_{k=1}^{K} p_\lambda(\theta_k), \tag{12}$$

where $p_\lambda(\theta)$ is the SCAD penalty function defined as

$$p_\lambda(\theta) = \begin{cases} \lambda|\theta| & \text{if } |\theta| \leq \lambda, \\ -\frac{|\theta|^2 - 2a\lambda|\theta| + \lambda^2}{2(a-1)} & \text{if } \lambda < |\theta| \leq a\lambda, \\ \frac{(a+1)\lambda^2}{2} & \text{if } |\theta| > a\lambda, \end{cases}$$

where $a > 2$ is a fixed parameter, typically set to 3.7 based on empirical evidence and λ controls the penalty's intensity. Unlike ridge, SCAD allows some coefficients to be exactly zero and, unlike lasso, the larger coefficients are shrunk less severely. A drawback of the SCAD penalty is that it is non-convex, which makes computation more difficult. We adopt the coordinate descent algorithm for non-convex penalized regression proposed in Breheny and Huang (2011) who provide a fast, efficient and stable algorithm for solving (12).

3.2.2 Bayesian variable selection

From a Bayesian point of view, variable selection is accomplished by specifying a prior distribution over the model parameters. A Bayesian counterpart to (5) can be written as

$$\dot{r}_{b,t} = -\theta^\intercal \dot{r}_{c,t} + \epsilon_t, \qquad \epsilon_t \sim N(0, \sigma_\epsilon^2), \qquad \theta \sim p(\theta|\omega), \tag{13}$$

where $p(\theta|\omega)$ defines the prior distribution over θ, and ω collects all hyperparameters. It is interesting to note that the maximum a posteriori probability estimator, obtained by minimizing the negative log-posterior, is equivalent to classical approaches of penalized regression, and specific choices of prior distributions can generate all classical penalized regressions presented above (see, for example, Bhadra, Datta, Polson, and Willard, 2019). A successful global-local shrinkage prior is the horseshoe (Carvalho, Polson, and Scott, 2010), which can be defined as

$$\theta \sim p(\theta|\lambda_i, \tau) \sim N(0, \lambda_i^2), \quad \lambda_i|\tau \sim C^+(0, \tau), \quad \tau \sim C^+(0, 1), \tag{14}$$

where C^+ denotes the half-Cauchy density, whose support is the non-negative real line.

The main advantages of the horseshoe prior are its tail- and sparse-robustness properties. These properties are a byproduct of its spike at zero and the heavy tails produced by the scale mixture of normals defined by (14), which shrink weak signals toward the origin while still allowing strong signals to remain unshrunk due to their heavy tails.

3.2.3 Ensembles

The ensemble approach to variable selection consists of integrating multiple models to identify and select the most significant variables in a given data set. Techniques such as *bagging* (Breiman, 1996) and *boosting* (Freund and Schapire, 1996; Zhang and Yu, 2005) are commonly used, where models are trained on various subsets or reweighted instances of the data. In this paper, we consider the $L2$-boosting method of Bühlmann and Yu (2003) which falls under the umbrella of boosting algorithms. The $L2$-boosting method is based on the principle of improving a model's predictive power by sequentially adding weak learners across a maximum of m^* iterations, where m^* is a hyperparameter that must be tuned. When the number of covariates K in a data set is large (and when selecting a small number of relevant covariates is desirable), boosting is usually superior to standard estimation techniques for regression models (such as backward stepwise linear regression, which, for example, cannot be applied if K is larger than the number of observations T). Algorithm 1 provides a pseudocode to implement the $L2$-boosting for minimum-variance parametric portfolios.[7]

[7] In Algorithm 1, $\hat{\theta}^{[m]}$ is the estimated value of θ obtained *during* the m-th iteration of the boosting algorithm, and $\hat{\theta}_m$ is the estimated value of θ obtained *after* m iterations of the boosting algorithm.

3.2.4 Choice of hyperparameters

In each estimation window, the regularization hyperparameters for ridge, lasso, elastic net, and SCAD are determined through five-fold cross-validation, as discussed in Hastie et al. (2009, Chapter 7). This involves setting a range of potential values for the hyperparameters. The sample is segmented into five segments, referred to as "folds." For each j in the range from 1 to 5, the jth fold is excluded, and the remaining four are used to generate predictions for each hyperparameter value. The prediction error, or cross-validation error, for each hyperparameter value is then calculated on the excluded jth fold. This procedure is repeated across all five folds. The hyperparameter value that results in the lowest average cross-validation error is chosen. As for the $L2$-boosting method, the hyperparameter of interest is the maximum number of boosting iterations, which is analogous to an early stopping criterion. Stopping the algorithm too early will not capture important features of the data. Terminating the process too late can lead to the well-known issue of overfitting. We follow Hurvich et al. (1998) and Schmid and Hothorn (2008) and employ the corrected AIC criterion to select the optimal number of iterations, m^*,

$$AIC(m) = \log\left(\hat{\text{var}}\left[r_{p,t+1}^{m}\right]\right) + \frac{1 + df_m/t}{1 - (df_m + 2)/t}, \tag{15}$$

where df is the effective number of degrees of freedom after m iterations of the boosting algorithm.[8] As m increases, $\hat{\text{var}}\left[r_{p,t+1}^{m}\right]$ decreases, and df_m increases. Therefore, the optimal number of iterations, m^*, is the one that corresponds to the smallest AIC. Finally, we set the step size ν to a fixed value of 0.10.

Finally, we perform full hierarchical Bayes estimation of the Bayesian regression model in (13) and (14) and learn about the hyperparameter τ in each estimation window. Estimation is performed using the efficient algorithm proposed by Bhattacharya, Chakraborty, and Mallick (2016), and the global-local parameters τ and λ_i are updated using the slice sampler of Polson, Scott, and Windle (2014).

[8] We also experimented with the five-fold cross-validation method in order to tune the maximum number of boosting iterations for the $L2$-boosting method. The results in terms of portfolio performance are slightly worse than those obtained with the corrected AIC criterion.

3.3 Interpreting predictor importance

We are also interested in understanding how and why a selected predictor is important from a minimum-variance portfolio construction perspective. Drawing on the approach developed in DeMiguel et al. (2020), we rewrite the minimum-variance optimization problem in eq. (3) as a quadratic optimization problem,

$$\min_{\theta} \ (1/2)\theta^{\intercal}\hat{\Sigma}_c\theta + \theta^{\intercal}\hat{\sigma}_{bc}, \tag{16}$$

where $\hat{\Sigma}_c$ is the sample covariance matrix of the predictor return vector r_c, and $\hat{\sigma}_{bc}$ is the sample vector of covariances between the benchmark portfolio return r_b and the predictor-return vector r_c. Eq. (16) decomposes the minimum-variance utility function into the variance of the predictor return vector, $(1/2)\theta^{\intercal}\hat{\Sigma}_c\theta$, and the covariance of the predictor returns with the benchmark returns, $\theta^{\intercal}\hat{\sigma}_{bc}$. This decomposition reveals that the utility achieved by a given method is fundamentally determined by two complementary aspects: i) the ability to select predictors that contribute to decrease the variance-covariance of the predictor return vector, and ii) the ability to select predictors that contribute to decrease the covariance of the predictor return vector with the benchmark return vector.

Eq. (16) allows us to evaluate the importance of a *individual* predictor. For that purpose, we calculate their *marginal contributions* using the first-order derivative of eq. (16),

$$\begin{aligned} \text{Predictor importance} &= \frac{\partial}{\partial\theta}\left(\frac{1}{2}\theta^{\intercal}\hat{\Sigma}_c\theta + \theta^{\intercal}\hat{\sigma}_{bc}\right) \qquad (17) \\ &= \underbrace{\mathrm{diag}(\hat{\Sigma}_c)\theta}_{\text{own-var.(pred.)}} + \underbrace{\hat{\Sigma}_c - \mathrm{diag}(\hat{\Sigma}_c)\theta}_{\text{cov.(pred.)}} + \underbrace{\hat{\sigma}_{bc}.}_{\text{cov.(bench.)}} \end{aligned}$$

The ith, $i = 1, \ldots, K$, component of the right-hand side in (17) is the marginal contribution of the ith term to the parametric portfolio minimum-variance utility; that is, the marginal change in minimum-variance utility associated with a unit increase in the weight that the parametric portfolio assigns to the ith term. The marginal contribution of an individual predictor can be decomposed into three terms: the predictor's own variance, $\mathrm{diag}(\hat{\Sigma}_c)\theta$, the predictor's covariance with other predictors, $(\hat{\Sigma}_c - \mathrm{diag}(\hat{\Sigma}_c))\theta$, and the covariance between the predictor and benchmark portfolios, $\hat{\sigma}_{bc}$.

4 Empirical application

We now perform an empirical analysis of the performance of minimum-variance parametric portfolios when employing alternative regularization and variable selection methods discussed in Section 3.2. The parametric portfolios are obtained by assuming an equally weighted (EW) benchmark portfolio. Two reasons motivate our choice of this benchmark portfolio. First, the well-documented good performance of this strategy relative to more sophisticated ones, as extensively discussed in DeMiguel et al. (2009). Second, Lassance and Martín-Utrera (2020) shows that the EW strategy emerges naturally in a mean-variance context since an optimal mean-variance portfolio can be viewed as a linear combination of the EW portfolio and an arbitrage portfolio.[9]

We consider two alternative configurations of the data set discussed in Section 2. First, we implement parametric portfolios using only the original 95 predictors mentioned in Gu et al. (2020, 2021). Second, we utilize the augmented set of 4,610 predictors, which includes second-order and cross-product transformations. This enables us to understand the extent to which non-linear transformations of firm characteristics improve portfolio performance.

Given our interest in the minimum-variance strategy, the most relevant performance metric is the out-of-sample ex-post (i.e. realized) standard deviation of out-of-sample portfolio returns. However, it is also common in the literature to report performance in terms of risk-adjusted returns measured by the Sharpe ratio (SR) both before and after transaction costs. The out-of-sample evaluation works as follows. First, we choose a window over which to perform the estimation. The total number of monthly observations in the dataset is $T_{tot} = 594$ and we choose an estimation window of $T = 120$. Second, using the return data over the estimation window, we compute the minimum-variance parametric portfolios using the ML methods detailed in Section 3.2. Third, we repeat this rolling-window procedure for the next month by including the data for the next month and dropping the data for the earliest month. We continue doing this until the end of the dataset is reached. At the end of this process, we have generated $T_{tot} - T = 474$ portfolio weight vectors, w_t^j, for $t = T, \dots, T_{tot} - 1$ and for each strategy j. Holding the portfolio w_t^j for one month gives the out-of-sample return net of transaction costs

[9] We obtain qualitatively similar results when assuming a value-weighted benchmark portfolio.

at time $t+1$:

$$r_{t+1}^{j} = \left(w_t^j\right)^{\top} r_{t+1} - c \times |w_t^j - \left(w_{t-1}^j\right)^+|,$$

where $|w_t^j - \left(w_{t-1}^j\right)^+|$ denotes the monthly portfolio turnover, c is the level of transaction costs, and $\left(w_{t-1}^j\right)^+$ is the portfolio for the jth strategy before rebalancing at time t, that is

$$\left(w_{t-1}^j\right)^+ = w_{t-1}^j \circ \left(e_{t-1} + r_t\right),$$

where e_{t-1} is the N_{t-1} dimensional vector of ones and $x \circ y$ is the element-wise product of vectors x and y. Then, for each portfolio we study, we compute the annualized out-of-sample standard deviation and the SR of returns net of transaction costs:

$$\begin{aligned}\hat{\sigma}^j &= \left(\frac{12}{T_{tot} - T} \sum_{t=T}^{T_{\text{tot}}-1} \left(\left(w_t^j\right)^{\top} r_{t+1} - \hat{\mu}^j\right)^2\right)^{1/2}, \\ \widehat{\mathrm{SR}}^j &= \frac{\hat{\mu}^j - Rf_t}{\hat{\sigma}^j},\end{aligned}$$

where $\hat{\mu}^j = \frac{12}{T_{\text{tot}} - T} \sum_{t=T}^{T_{\text{tot}}-1} \left(w_t^j\right)^{\top} r_{t+1}$ and Rf_t denote the risk-free rate at time t.[10]

We consider two levels of transaction costs: 0 basis points (b.p.) and 10 b.p. We also test for the statistical significance of the differences in the portfolio variances and the Sharpe Ratio (SR) of the two portfolios by using the two-sided p-value of the prewhitened HAC_{PW} test described by Ledoit and Wolf (2011) and Ledoit and Wolf (2008) for the portfolio variance and the SR, respectively. Specifically, we test for differences in portfolio variances and SR of the various strategies in relation to those of the minimum-variance parametric portfolio obtained when all 4,610 predictors are used in conjunction with the lasso variable selection method.

Competing strategies. We also implement alternative portfolio strategies for comparison purposes. First, we implement the classic equally weighted (EW) and value-weighted (VW) strategies. These two strategies do not require estimating model parameters and are easily scalable to large cross-sections of assets. In a parametric portfolio context, investing in these two strategies corresponds to an extreme regularization of the parameters, in which all coefficients are set to zero. Second, we implement the minimum variance parametric portfolios using the

[10] The risk free rate was obtained from Ken French's data library web site.

characteristics from the 3-factor (market, size, and value) and 5-factor (market, size, value, profitability, and investment) models of Fama and French (1993) and Fama and French (2015), respectively. Finally, we implement the minimum variance portfolios using the 95 characteristics along with the traditional OLS method to estimate the coefficients of the parametric policy.

4.1 Results

We report in Table 2 the out-of-sample standard deviation of minimum-variance portfolio returns. The Table allows us to draw several important conclusions. First and foremost, the choice of the variable selection approach matters: the different methods lead to minimum-variance parametric portfolios with varying performance. We observe that the lasso method results in minimum-variance portfolios with lower portfolio risk compared to all other methods for both configurations of the data set. Three methods emerge as the best performers in terms of the standard deviation of portfolio returns: lasso, elastic net, and $L2$-boosting. The best overall performance in terms of annualized standard deviation of portfolio returns is achieved when using the 4,610 predictors and the lasso method (8.3%). This figure, however, is statistically indistinguishable from that obtained with the elastic net (8.4%) and with the $L2$-boosting (8.8%) methods.

Second, all variable selection methods lead to minimum-variance parametric portfolios that outperform those obtained with ad-hoc selected characteristics, as well as portfolios obtained using the OLS method. Specifically, using the characteristics from the popular 3-factor and 5-factor models leads to a portfolio standard deviation of 15%, whereas using the OLS method leads to a portfolio standard deviation of 18%, and both figures are substantially higher relative to those obtained with ML-based portfolios. Furthermore, the classic EW and VW strategies also exhibit higher standard deviations than the ML-based portfolios.

Third, we observe a notable decrease in the standard deviation of portfolio returns performance when utilizing the expanded set of 4,610 predictors compared to the original 95 predictors. For the lasso method, the standard deviation decreases by approximately 9%, dropping from 9.1% to 8.3% – and the difference is statistically significant. Similarly, the adalasso method shows a decrease of about 5%, whereas the elastic net method also experiences a reduction of approximately 9%. The $L2$-boosting method exhibits a modest decrease in standard deviation of about 2% with the augmented data set, while the Bayesian horseshoe method yields similar results in both configurations of the data set.

It is interesting to note that the opposite result is obtained when using the ridge method: the portfolio standard deviation *increases* when using the augmented data set. The rationale for this result lies in the inherent features of the ridge approach: the dense estimation obtained with the ridge methods implies that all predictors are Incorporated into the model to some extent. With such a large number of predictors, the dense estimation demands very strong regularization to properly estimate the model. As a result, all coefficients are shrunk to very small values, and the ridge method with the augmented data set begins to resemble that of the classic EW strategy, which is the benchmark portfolio policy used in the parametric portfolio; see eq. (1). Not surprisingly, the standard deviation of the ridge-based portfolios with the augmented data set is close to that obtained with the EW and VW strategies. This outcome highlights a limitation of the ridge method in scenarios where the predictor set is vastly expanded relative to the number of data points.

Table 3 reports the annualized Sharpe ratios both before and after transaction costs of 10 basis points, along with the average monthly turnover of all strategies. We observe that portfolios based on the lasso, elastic net, and SCAD methods perform poorly in terms of Sharpe ratio, both before and after transaction costs. For example, the after-fee Sharpe ratio obtained with the lasso method varies between -0.12 and -0.32 depending on the data set configuration. When transaction costs are taken into account, only the $L2$-boosting and ridge methods yield parametric portfolios with positive Sharpe ratios (1.076 and 0.814 when using the 95 characteristics).[11] The $L2$-boosting portfolio is the best performer in terms of before- and after-fee Sharpe ratios among all competing strategies. We also observe that the Sharpe ratios obtained with the augmented data set are often *worse* than those obtained with the restricted data set, indicating that increasing the predictor space can negatively affect performance in terms of risk-adjusted returns due to lower average portfolio returns.

The results in Table 3 reveal a much larger dispersion in risk-adjusted performance across variable selection methods relative to the dispersion observed in the standard deviation of portfolio returns, as shown in Table 2. This suggests that differences in the Sharpe ratio are

[11] Similar to the findings in Table 2, the performance of ridge-based portfolios in terms of risk-adjusted returns is comparable to that obtained with the equally weighted portfolio strategy.

driven mostly by the *average* portfolio returns rather than the *standard deviation* of portfolio returns. Moreover, Table 3 indicates an apparent superiority of the $L2$-boosting method. To help understand the differences in risk-adjusted performance across methods, eq. (2) shows that the differences in parametric portfolio returns are fully determined by differences in $\theta^\intercal r_c$, since the term r_b is the same for all parametric portfolio strategies. In other words, the predictors selected by each method and how each method invests in the selected predictors are the main drivers of performance in terms of returns.

Figure 1 plots the average values of the aggregate monthly returns for the predictors selected with each method. Specifically, the figure displays the average values of $\theta^\intercal r_c$ calculated across all rolling estimation windows. Consistent with the results reported in Table 3, minimum-variance portfolios obtained with the $L2$-boosting method achieve the highest value (1% per month), whereas the SCAD method achieves the lowest (-1% per month). As expected, the predictor return obtained with the ridge method is very close to zero, as its performance resembles that of the benchmark strategy.

One important limitation of Figure 1 is the lack of information about which *individual* predictor contributes the most to performance in terms of (aggregate) predictor returns. To complement the results in Figure 1, we plot in Figure 2 the returns on the top-10 and bottom-10 individual predictors for each method. We observe that the interaction between size ($mve0$) and zero trading days ($zerotrade$) contributes the most to the positive risk-adjusted performance obtained with the $L2$-boosting method (0.22% per month). For the other methods, the negative returns among the bottom-10 predictors are often larger in magnitude relative to those of the top-10 predictors, which helps explain not only the negative performance in terms of aggregate predictor return displayed in Figure 2, but also the negative performance in terms of risk-adjusted returns reported in Table 3.

The fourth column of Table 3 reports the average monthly portfolio turnover. We observe that portfolio turnover is significantly influenced by the choice of variable selection methods. Lasso, elastic net, and $L2$-boosting achieve portfolio turnovers around 1.5, whereas ridge-based portfolios have a much lower turnover of around 0.2. All variable selection methods yield portfolios with much lower turnover relative to that obtained with OLS-based portfolios (5.14). Ad-hoc selected characteristic-based portfolios and traditional benchmark strategies, such as

EW and VW portfolios, typically have lower turnover compared to ML methods. This is because these strategies are based on simpler allocation rules that do not change frequently with market conditions and, consequently, require less portfolio rebalancing.

We also report in Table 4 the performance of the various portfolio strategies in terms of maximum drawdown and value-at-risk (VaR) based on the historical simulation method with 99% confidence. The Table shows that adalasso and $L2$-boosting methods outperformed all other benchmarks in terms of both maximum drawdown and VaR. Out of all the methods, $L2$-boosting demonstrated the most superior performance, with a maximum drawdown of 38.6% and a VaR of 7.4%. The OLS method performed poorly compared to the other methods, with a maximum drawdown of 68.9% and a VaR of 14.8%. These results show that ML-based portfolios are also effective in minimizing tail risks.

In summary, the results reported in Tables 2 to 4 reveal that selecting predictors with variable selection methods leads to minimum-variance portfolios with superior performance relative to traditional OLS methods and ad-hoc characteristic-based portfolios, in terms of lower risk, smaller maximum drawdowns, and VaR. Moreover, the results obtained with the $L2$-boosting method showed that lower portfolio risk can be achieved without compromising risk-adjusted returns: the performance measured by the Sharpe ratio is substantially higher relative to all other competing strategies, even when transaction costs are taken into account.

4.2 Predictor importance

To further understand the aspects that contribute to the performance of minimum-variance portfolios, we study the profile of selected predictors as well as their distribution across predictor classes (main effects, second-order effects, and interaction effects.). The results reported in Table 5 reveal distinct patterns in predictor selection across different methods. Specifically, the average number of selected predictors varies from 16 (SCAD) to 75 ($L2$-boosting). The best performing method in delivering portfolios with lower risk (lasso) selects 33 predictors on average. This result suggests that the number of selected predictors vastly exceeds the number of variables used in popular factor models, suggesting that ad-hoc sparsity can be detrimental to portfolio performance.

The "class-relative percentage" panel in Table 5 reports the distribution of the selected predictors within each class (main effects, second-order effects, and interaction effects) for each method. Across all methods, interaction effects consistently account for the highest percentage of selected predictors. This finding is not surprising, given the composition of the predictor space: with a total of 4610 predictors, only 95 (2.1%) are main effects, 88 (1.9%) are second-order effects, and the remaining 4427 (96%) are interaction terms. Thus, interaction effects naturally represent the largest pool of potential predictors and are expected to make up the majority of selected predictors simply due to their abundant quantity. Despite this expectation, the distribution across classes varies by method, reflecting each method's tendency to prioritize different classes of predictors. For example, SCAD places a much higher emphasis on main effects (16%) compared to other methods, such as adalasso (0.8%) and $L2$-boosting (1.4%). In contrast, methods like adalasso and elastic net devote a higher percentage to interaction effects.

The "difference from equal-importance" panel in Table 5 shows how each method's selection deviates from a benchmark based on equal-importance weighting. If the classes were equally important, one would expect the selection of predictors to be proportional to the number of predictors available in each class. Therefore, deviations from this equal-importance benchmark reveal which classes are being over- or under-selected relative to what would be expected by chance. We observe that lasso, elastic net, horseshoe and SCAD *underrepresent* interaction effects compared to an equal-importance distribution, as shown by the *negative* deviations for this class, and *overrepresent* main effects and second-order effects, as shown by the *positive* deviations for these classes.

We also investigate the importance of individual predictors using the method described in Section 3.3. Figures 3, 4, and 5 show the marginal contributions of the top-20 selected predictors for lasso, elastic net, and $L2$-boosting, respectively.[12] Each Figure breaks down the contributions into three components: the predictor's own variance, its covariance with other predictors, and its covariance with the benchmark portfolios (see Eq. (17)). Predictor importance based on own-variance (left-hand plot) is ranked in ascending order: predictors whose own variance leads to smaller increases in the objective function (Eq. (16)) are more important. Predictor importance based on the correlation with other predictors (center plot) and on the correlation with the benchmark portfolio (right-hand plot) is ranked in descending

[12] Marginal contributions are calculated across the out-of-sample period, and the reported values are averages across all estimation rounds.

order: predictors with lower covariance with other predictors and with the benchmark portfolio contribute more to decrease the objective function and are thus more important. Finally, predictors with a positive (negative) parametric-portfolio coefficient are in blue (red).

Figures 3-5 show that the importance of individual predictors varies across methods. Some predictors like market beta, return volatility (*retvol*), idiosyncratic volatility (*idiovol*), and maximum daily return (*maxret*) are selected by lasso and elastic net mainly for their ability to improve investors' utility because of their covariance with the benchmark portfolio. These results align with those reported in Scherer (2011), who argue that risk-based characteristics are important drivers of the minimum-variance portfolio allocations. These four predictors have a negative parametric-portfolio coefficient, indicating that stocks with higher values on those predictors are assigned lower weights in the minimum-variance portfolio. The lasso method also prioritizes stocks with higher values of size (*mvel1*) and short-term reversal (*mom1m*). As for the $L2$-boosting method, square values of one-year momentum (*mom12m_sq*) are important because they covary with the benchmark portfolio, whereas square values of short-term reversals (*mom*1*m_sq*) matter because they have low covariance with the other predictors.

It is also remarkable that interactions between risk-based and liquidity measures are among the most important predictors for the three methods. For instance, interactions between market beta and standard deviation of liquidity (*beta*×*std_dolvol*), liquidity and idiosyncratic volatility (*dolvol* × *idiovol*), and bid-ask spread and liquidity (*baspread* × *dolvol*) are among the most important predictors for the three methods. To understand how these three interactions connect to optimal portfolio allocations, we plot in Figures 6 to 8 the average minimum-variance portfolio weight across standardized values of the first characteristic and across quintiles of the second characteristic. We find that stocks with lower liquidity get higher weights, and this result is stronger for stocks with lower idiosyncratic volatility. Stocks with low market betas get higher weights, and this result is stronger for stocks with higher standard deviation of liquidity. Finally, we observe a pronounced u-shaped non-linear relation between bid-ask spreads and portfolio weights, as well as a pronounced interaction with volatility-based predictors.

The analysis of predictor importance shown in Figures 3-5 allows us to draw three main conclusions. First, we show that risk-based predictors are not the only class of important predictors for constructing minimum-variance portfolios. Our results reveal that size-, momentum-, and liquidity-based predictors are also among the most important predictors via their main effects as well as non-linear transformations (especially interactions). Second, we observe that some predictors that help decrease portfolio risk can also help increase portfolio returns in some situations. For instance, in the case of the $L2$-boosting method, Figure 16 shows that the interaction between size and zero trading days appears among the most important predictors that help decrease portfolio risk via its lower correlation with the benchmark portfolio. This predictor also contributes to *increasing* portfolio *returns*, as reported in Figure 2.

We also study the economic significance of the most important risk-based and liquidity predictors as well as their interactions. Table 6 presents the regression results of selected parametric-portfolio predictor coefficients on three state variables: investor sentiment, market liquidity, and economic recession. Panel A shows the main effects of each predictor, while Panel B reports the interaction effects. For the predictor *beta* the results indicate a significant relationship with investor sentiment across all methods, with consistently negative coefficients. Additionally, the variations in the *beta* parametric-portfolio coefficient is also significantly associated with the recession indicator in the $L2$-boosting method, where it shows a negative coefficient.

Moreover, some of the interaction effects show significant relationships with state variables even when the corresponding main effects are not significant. This suggests that interactions may capture more complex or nuanced dynamics that are not apparent in the main effects alone. For instance, The *dolvol_x_idiovol* interaction also exhibits significant relationships with investor sentiment across several methods, all with positive coefficients. The liquidity variable has significant effects on the *beta_x_idiovol* interaction the $L2$-boosting method, showing a negative coefficient.

Finally, we also examine the temporal dynamics of the estimated coefficients of the selected characteristics with the lasso, elastic net, and $L2$-boosting methods across all estimation rounds. For that purpose, we plot in Figures 9 to 17 a heatmap of the estimated coefficients for the features' main effect, second-order effect, and interactions, respectively. The Figures reveal that the selection of relevant characteristics and the strength of their effects exhibit substantial variations over time. Many characteristics appear and disappear from the chosen subset, suggesting that the importance of firm characteristics for portfolio choice is highly time-varying.

4.3 Are the portfolios implementable?

To assess the practicality of implementing the ML-based minimum-variance portfolios, Table 7 reports descriptive statistics of the distribution of portfolio weights across all estimation rounds. The Table shows that the minimum and maximum weights of the ML-based portfolios are not extreme, ranging from -0.8% to 0.8%. These values are well within the acceptable range for traditional investment strategies. Moreover, the proportion of negative weights (short proportion) is also relatively low, hovering around 30%. This suggests that the ML-based portfolios are long-biased and do not involve excessive short selling.

It is interesting to note that, according to Green and Hollifield (1992), the presence of a dominant first principal component (or factor) would result in extreme negative weights in minimum-variance portfolios. Our results seem to contradict this finding, as they indicate the absence of extreme weighting. One plausible explanation for our results is that our methodology involves constructing a minimum-variance portfolio using a *multitude* of characteristics-based factors. In this situation, DeMiguel et al. (2020) and DeMiguel, Martin-Utrera, and Uppal (2024) show that exploiting multiple factors leads to trading diversification (i.e. netting of trades across factors), which helps reduce extreme positions in individual stocks.

4.4 Screening predictors

The results reported in Table 3 show that selecting a few predictors ad hoc, as in the popular 3-factor and 5-factor models, leads to minimum-variance parametric portfolios with poor performance relative to those obtained with ML variable selection methods. To avoid selecting predictors ad hoc, we implement the marginal screening (MS) method of Fan and Lv (2008) which employs a data-driven approach to screen a vast pool of predictors based on their individual correlations with the target variable. MS is computationally efficient, making it well-suited for ultrahigh dimensions. Moreover, MS ensures the exact recovery of true non-zero coefficients under specific sparsity and signal strength conditions (see Genovese, Jin, Wasserman, and Yao, 2012). Algorithm 2 describes the procedure for implementing the

MS method.

We implement the MS method to screen the augmented set of 4,610 features to select the best 3, 5, 10, and 50 predictors. We then use the selected predictors to construct minimum-variance portfolios. The results reported in Table 8 reveal that selecting only the top-3 predictors using MS leads to minimum-variance portfolio returns with an annualized standard deviation of 10.7%, which is substantially *lower* relative to those obtained when using ad-hoc selected characteristics, as reported in Table 3. The performance of the MS methods peaks when the top-10 predictors are selected (9.5%). However, the performance of screening-based portfolios is worse than those obtained with the variable selection methods, as reported in Table 3. We conclude from this analysis that although MS is a much better alternative in comparison to selecting predictors ad-hoc, the use of more sophisticated variable selection methods can offer further improvements.

5 Concluding remarks

This paper provides evidence that machine learning (ML) can be a powerful tool for selecting variables relevant for optimal portfolio choice. We focus on directly parameterizing portfolio weights as a function of lagged firm-level predictors. By employing ML variable selection methods on a large pool of features and their second-order and cross-product transformations (4,610 predictors), we find that ML can identify a subset of important variables that outperform traditional low-dimensional factor models in terms of lower risk and higher risk-adjusted returns. We also find that the choice of variable selection method matters, with different methods selecting different subsets of features. The $L2$-boosting method emerged as the most comprehensive approach, as it was able to identify predictors that minimize risk, reduce covariance with the benchmark, and increase risk-adjusted portfolio returns.

Our findings have several implications for practitioners. First, they suggest that ML can be a valuable tool for improving portfolio performance. Second, they highlight the importance of accounting for the interaction between variables in the portfolio construction process. Third, they indicate that the number of selected features relevant to the portfolio construction problem is greater than the number of variables used in popular factor models. As a result, practitioners should consider using ML variable selection methods to identify a subset of important variables for their portfolios.

References

Aït-Sahalia, Y. and M. W. Brandt (2001). Variable selection for portfolio choice. *The Journal of Finance 56*(4), 1297–1351.

Ang, A., A. Madhavan, and A. Sobczyk (2017). Crowding, capacity, and valuation of minimum volatility strategies. *The Journal of Index Investing 7*(4), 41.

Bai, J. and S. Ng (2009). Boosting diffusion indices. *Journal of Applied Econometrics 24*(4), 607–629.

Baker, M. and J. Wurgler (2006). Investor sentiment and the cross-section of stock returns. *The journal of Finance 61*(4), 1645–1680.

Bhadra, A., J. Datta, N. G. Polson, and B. Willard (2019). Lasso meets horseshoe. *Statistical Science 34*(3), 405–427.

Bhattacharya, A., A. Chakraborty, and B. K. Mallick (2016). Fast sampling with gaussian scale mixture priors in high-dimensional regression. *Biometrika*, asw042.

Brandt, M. W., P. Santa-Clara, and R. Valkanov (2009). Parametric portfolio policies: Exploiting characteristics in the cross-section of equity returns. *The Review of Financial Studies 22*(9), 3411–3447.

Breheny, P. and J. Huang (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Annals of Applied Statistics 5*(1), 232–253.

Breiman, L. (1996). Bagging predictors. *Machine learning 24*, 123–140.

Britten-Jones, M. (1999). The sampling error in estimates of mean-variance efficient portfolio weights. *The Journal of Finance 54*(2), 655–671.

Bryzgalova, S., M. Pelger, and J. Zhu (2020). Forest through the trees: Building cross-sections of stock returns. *Available at SSRN 3493458*.

Buchen, T. and K. Wohlrabe (2011). Forecasting with many predictors: Is boosting a viable alternative? *Economics Letters 113*(1), 16–18.

Bühlmann, P. and B. Yu (2003). Boosting with the L_2 loss. *Journal of the American Statistical Association 98*(462), 324–339.

Carvalho, C. M., N. G. Polson, and J. G. Scott (2010). The horseshoe estimator for sparse signals. *Biometrika 97*(2), 465–480.

Chan, L. K., J. Karceski, and J. Lakonishok (1999). On portfolio optimization: Forecasting covariances and choosing the risk model. *The review of Financial studies 12*(5), 937–974.

Chen, L., M. Pelger, and J. Zhu (2020). Deep learning in asset pricing. *Available at SSRN 3350138*.

Chinco, A., A. D. Clark-Joseph, and M. Ye (2019). Sparse signals in the cross-section of returns. *The Journal of Finance 74*(1), 449–492.

Clarke, R. G., H. De Silva, and S. Thorley (2006). Minimum-variance portfolios in the us equity market. *The Journal of Portfolio Management 33*(1), 10–24.

Clarke, R. G., H. De Silva, and S. Thorley (2011). Minimum-variance portfolio composition. *Journal of Portfolio Management 37*(2), 31.

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

DeMiguel, V., L. Garlappi, and R. Uppal (2009). Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *The review of Financial studies 22*(5), 1915–1953.

DeMiguel, V., J. Gil-Bazo, F. J. Nogales, and A. A. Santos (2023). Machine learning and fund characteristics help to select mutual funds with positive alpha. *Journal of Financial Economics 150*(3), 103737.

DeMiguel, V., A. Martin-Utrera, F. J. Nogales, and R. Uppal (2020). A transaction-cost perspective on the multitude of firm characteristics. *The Review of Financial Studies 33*(5), 2180–2222.

DeMiguel, V., A. Martin-Utrera, and R. Uppal (2024). Can competition increase profits in factor investing? *forthcoming in Management Science.*

DeMiguel, V. and F. J. Nogales (2009). Portfolio selection with robust estimation. *Operations research 57*(3), 560–577.

Dong, X., Y. Li, D. E. Rapach, and G. Zhou (2022). Anomalies and the expected market return. *The Journal of Finance 77*(1), 639–681.

Döpke, J., U. Fritsche, and C. Pierdzioch (2017). Predicting recessions with boosted regression trees. *International Journal of Forecasting 33*(4), 745–759.

Fabozzi, F. J., D. Huang, F. Jiang, and J. Wang (2024). What difference do new factor models make in portfolio allocation? *Journal of International Money and Finance 140.*

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

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

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association 96*(456), 1348–1360.

Fan, J. and J. Lv (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society Series B: Statistical Methodology 70*(5), 849–911.

Freund, Y. and R. E. Schapire (1996). Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on Machine Learning.* Citeseer.

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

Genovese, C. R., J. Jin, L. Wasserman, and Z. Yao (2012). A comparison of the lasso and marginal regression. *The Journal of Machine Learning Research 13*(1), 2107–2143.

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

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

Gu, S., B. Kelly, and D. Xiu (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies 33*(5), 2223–2273.

Gu, S., B. Kelly, and D. Xiu (2021). Autoencoder asset pricing models. *Journal of Econometrics 222*(1), 429–450.

Harvey, C. R., Y. Liu, and H. Zhu (2016). . . . and the cross-section of expected returns. *The Review of Financial Studies 29*(1), 5–68.

Harvey, C. R. and A. Siddique (2000). Conditional skewness in asset pricing tests. *The Journal of finance 55*(3), 1263–1295.

Hastie, T., R. Tibshirani, J. H. Friedman, and J. H. Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction*, Volume 2. Springer.

Hoerl, A. E. and R. W. Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 12*(1), 55–67.

Hong, H., T. Lim, and J. C. Stein (2000). Bad news travels slowly: Size, analyst coverage, and the profitability of momentum strategies. *The Journal of Finance 55*(1), 265–295.

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

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

Hurvich, C. M., J. S. Simonoff, and C.-L. Tsai (1998). Smoothing parameter selection in nonparametric regression using an improved akaike information criterion. *Journal of the Royal Statistical Society. Series B (Statistical Methodology) 60*(2), 271–293.

Jagannathan, R. and T. Ma (2003). Risk reduction in large portfolios: Why imposing the wrong constraints helps. *The journal of finance 58*(4), 1651–1683.

Kaniel, R., Z. Lin, M. Pelger, and S. Van Nieuwerburgh (2023). Machine-learning the skill of mutual fund managers. *Journal of Financial Economics 150*(1), 94–138.

Kauppi, H. and T. Virtanen (2021). Boosting nonlinear predictability of macroeconomic time series. *International Journal of Forecasting 37*(1), 151–170.

Kelly, B. and D. Xiu (2023). Financial machine learning. *Becker-Friedman Institute Working Paper 2023-100*.

Kempf, A. and C. Memmel (2006). Estimating the global minimum variance portfolio. *Schmalenbach Business Review 58*(4), 332–348.

Kozak, S., S. Nagel, and S. Santosh (2020). Shrinking the cross-section. *Journal of Financial Economics 135*(2), 271–292.

Lamoureux, C. G. and H. Zhang (2024). An empirical assessment of characteristics and optimal portfolios. *The Review of Asset Pricing Studies*, raae006.

Lassance, N. and A. Martín-Utrera (2020). Shrinking against sentiment: Exploiting behavioral biases in portfolio optimization. *Available at SSRN 3551224*.

Ledoit, O. and M. Wolf (2008). Robust performance hypothesis testing with the Sharpe ratio. *Journal of Empirical Finance 15*(5), 850–859.

Ledoit, O. and M. Wolf (2011). Robust performance hypothesis testing with the variance. *Wilmott Magazine 55*, 86–89.

Li, B. and A. G. Rossi (2020). Selecting mutual funds from the stocks they hold: A machine learning approach. *Available at SSRN 3737667*.

Masini, R. P., M. C. Medeiros, and E. F. Mendes (2023). Machine learning advances for time series forecasting. *Journal of economic surveys 37*(1), 76–111.

McLean, R. D. and J. Pontiff (2016). Does academic research destroy stock return predictability? *The Journal of Finance 71*(1), 5–32.

Medhat, M. and M. Schmeling (2022). Short-term momentum. *The Review of Financial Studies 35*(3), 1480–1526.

Moritz, B. and T. Zimmermann (2016). Tree-based conditional portfolio sorts: The relation between past and future stock returns. *Available at SSRN 2740751*.

Nagel, S. (2021). Machine learning in asset pricing. In *Machine Learning in Asset Pricing*. Princeton University Press.

Pástor, L. and R. F. Stambaugh (2000). Comparing asset pricing models: an investment perspective. *Journal of Financial Economics 56*(3), 335–381.

Pástor, L. and R. F. Stambaugh (2003). Liquidity risk and expected stock returns. *Journal of Political economy 111*(3), 642–685.

Polson, N. G., J. G. Scott, and J. Windle (2014). The bayesian bridge. *Journal of the Royal Statistical Society Series B: Statistical Methodology 76*(4), 713–733.

Rapach, D. E., J. K. Strauss, and G. Zhou (2013). International stock return predictability: what is the role of the united states? *The Journal of Finance 68*(4), 1633–1662.

Scherer, B. (2011). A note on the returns from minimum variance investing. *Journal of Empirical Finance 18*(4), 652–660.

Schmid, M. and T. Hothorn (2008). Boosting additive models using component-wise p-splines. *Computational Statistics & Data Analysis 53*(2), 298–311.

Simon, F., S. Weibels, and T. Zimmermann (2022). Deep parametric portfolio policies. *SSRN Working Paper 4150292*.

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

Zhang, T. and B. Yu (2005). Boosting with early stopping: Convergence and consistency. *Annals of Statistics*, 1538–1579.

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

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

Zou, H. and H. H. Zhang (2009). On the adaptive elastic-net with a diverging number of parameters. *Annals of statistics 37*(4), 1733.

Table 1: **Descriptive statistics of predictor returns**

The table reports descriptive statistics of predictor returns. The following statistics are reported: average monthly return, the standard deviation of monthly returns, the 25th and 75th percentiles, and the correlation of the predictor return with the equally weighted and value-weighted portfolio returns (ρ_{ew} and ρ_{vw}, respectively). The first four rows report aggregate values across i) all predictors, ii) predictors' main effects, iii) predictor's second-order effects, and iv) predictors' interaction effects. The remaining rows report individual statistics for the top and bottom 10 predictors within each category sorted in terms of average monthly return.

	Mean return (%)	Std. deviation (%)	25th (%)	75th (%)	ρ_{ew}	ρ_{vw}
All predictors	-0.003	0.523	-0.233	0.232	0.02	0.04
Main effects	-0.009	0.797	-0.367	0.352	0.04	0.06
Second-order effects	-0.024	0.706	-0.373	0.273	0.26	0.17
Interaction effects	-0.002	0.513	-0.227	0.229	0.02	0.04
Top-10 main effects						
indmom	0.309	1.545	-0.360	1.039	-0.12	-0.04
agr	0.275	0.653	-0.063	0.541	-0.21	-0.34
mom12m	0.270	1.863	-0.354	1.096	-0.33	-0.13
ill	0.205	1.206	-0.424	0.555	0.25	-0.07
mve0	0.186	0.917	-0.273	0.715	-0.57	-0.13
sp	0.154	0.783	-0.301	0.506	0.20	-0.01
mom6m	0.148	1.946	-0.399	0.941	-0.39	-0.21
rd_mve	0.139	0.658	-0.171	0.356	0.45	0.25
orgcap	0.105	0.668	-0.227	0.364	0.34	0.13
bm	0.102	0.639	-0.258	0.448	-0.21	-0.24
Bottom-10 main effects						
mom1m	-0.561	1.859	-1.125	0.211	-0.40	-0.25
maxret	-0.310	2.117	-1.353	0.479	0.69	0.43
turn	-0.248	1.533	-1.002	0.475	0.59	0.60
chmom	-0.248	1.289	-0.723	0.285	-0.33	-0.25
invest	-0.212	0.547	-0.494	0.111	0.24	0.35
retvol	-0.197	2.569	-1.472	0.708	0.73	0.47
lgr	-0.192	0.454	-0.417	0.033	0.44	0.41
chcsho	-0.179	0.506	-0.420	0.077	0.34	0.42
sgr	-0.172	0.542	-0.433	0.127	0.41	0.41
hire	-0.170	0.530	-0.413	0.118	0.27	0.36
Top-10 second-order effects						
ill_sq	0.208	0.985	-0.325	0.516	0.25	-0.04

Table 1 continued on next page

Table 1 continued from previous page

	Mean return (%)	Std. deviation (%)	25th (%)	75th (%)	ρ_{ew}	ρ_{vw}
mom12m_sq	0.170	1.506	-0.561	0.667	0.70	0.54
indmom_sq	0.132	1.303	-0.548	0.712	0.53	0.41
rd_mve_sq	0.114	0.539	-0.161	0.290	0.42	0.21
mom6m_sq	0.106	1.817	-0.695	0.644	0.67	0.49
sp_sq	0.087	0.610	-0.265	0.385	0.26	0.02
mve0_sq	0.087	0.618	-0.231	0.460	-0.61	-0.20
std_dolvol_sq	0.086	0.823	-0.307	0.561	-0.23	-0.44
orgcap_sq	0.072	0.551	-0.203	0.271	0.29	0.07
ps_sq	0.071	0.713	-0.193	0.416	-0.52	-0.26
Bottom-10 second-order effects						
maxret_sq	-0.322	1.749	-1.192	0.311	0.61	0.33
turn_sq	-0.257	1.175	-0.814	0.293	0.52	0.53
retvol_sq	-0.227	2.276	-1.366	0.539	0.66	0.39
agr_sq	-0.195	0.516	-0.457	0.072	0.61	0.52
invest_sq	-0.183	0.481	-0.437	0.062	0.47	0.40
chcsho_sq	-0.144	0.419	-0.339	0.081	0.29	0.37
lgr_sq	-0.139	0.415	-0.354	0.074	0.52	0.39
grltnoa_sq	-0.135	0.381	-0.323	0.053	0.39	0.35
sgr_sq	-0.130	0.538	-0.432	0.122	0.56	0.39
hire_sq	-0.127	0.465	-0.379	0.102	0.60	0.46
Top-10 interaction effects						
ill × mve0	0.953	0.988	0.272	1.417	0.09	-0.15
mom1m × mom6m	0.496	1.522	-0.150	0.888	0.31	0.19
mom12m × mom1m	0.364	1.405	-0.223	0.754	0.26	0.13
mve0 × std_turn	0.333	0.934	-0.120	0.759	-0.17	0.19
agr × maxret	0.326	0.764	-0.053	0.642	-0.31	-0.36
agr × retvol	0.318	0.798	-0.058	0.660	-0.34	-0.38
baspread × mve0	0.317	0.887	-0.124	0.750	-0.33	0.10
agr × baspread	0.292	0.785	-0.077	0.611	-0.32	-0.33
maxret × mom12m	0.291	1.869	-0.232	1.111	-0.30	-0.11
mve0 × retvol	0.287	0.917	-0.124	0.753	-0.31	0.12
Bottom-10 interaction effects						
mom1m × retvol	-0.562	1.890	-1.063	0.217	-0.35	-0.21
maxret × mom1m	-0.491	1.705	-0.944	0.165	-0.30	-0.18
baspread × mom1m	-0.478	1.907	-1.012	0.249	-0.33	-0.18

Table 1 continued on next page

Table 1 continued from previous page

	Mean return (%)	Std. deviation (%)	25th (%)	75th (%)	ρ_{ew}	ρ_{vw}
idiovol × mom1m	-0.472	1.763	-1.012	0.262	-0.37	-0.23
mom1m × tang	-0.443	1.396	-0.859	0.109	-0.37	-0.23
beta × mom1m	-0.440	1.798	-1.040	0.289	-0.38	-0.26
age × mom1m	-0.435	1.064	-0.810	0.061	-0.42	-0.27
dolvol × mom1m	-0.410	1.730	-0.899	0.251	-0.39	-0.26
mom1m × std_dolvol	-0.409	1.484	-0.834	0.186	-0.34	-0.20
mom1m × ps	-0.403	1.173	-0.779	0.105	-0.40	-0.27

Table 2: **Standard deviation of out-of-sample minimum-variance portfolio returns**

The table reports annualized out-of-sample (OOS) standard deviation of portfolio returns for the minimum-variance parametric portfolio strategy when the ML methods discussed in Section 3.2 are used to perform variable selection and regularization. The table also reports the performance of the minimum-variance strategy when using only the market, size, value, profitability, and investment characteristics as well as the performance of the equally weighted and value-weighted strategies. The ML-based portfolios are implemented with two alternative configurations of the data set discussed in Section 2: using only the original 95 predictors used in Gu et al. (2020, 2021) and ii) all 4,610 predictors including the second-order and cross-products transformations. One, two, and three asterisks indicate that the differences in portfolio variance with respect to those of the lasso strategy with 4,610 predictors are significant at the 10%, 5% and 1% level, respectively.

	Std. dev. of OOS portfolio returns
market, size, value	0.150***
market, size, value, profit., invest.	0.146***
equally weighted	0.183***
value-weighted	0.158***
OLS	0.176***
Lasso	
(95 characteristics)	0.091*
(4,610 characteristics)	0.083
Adalasso	
(95 characteristics)	0.097**
(4,610 characteristics)	0.091**
Elastic net	
(95 characteristics)	0.092
(4,610 characteristics)	0.084
Ridge	
(95 characteristics)	0.117***
(4,610 characteristics)	0.158***
Horseshoe	
(95 characteristics)	0.089*
(4,610 characteristics)	0.094**
Scad	
(95 characteristics)	0.100***
(4,610 characteristics)	0.091**
$L2$-boosting	
(95 characteristics)	0.090**
(4,610 characteristics)	0.087

Table 3: **Out-of-sample performance statistics of various portfolio policies**

The table reports annualized out-of-sample risk-adjusted portfolio returns measured by the Sharpe ratio (SR). Minimum-variance parametric portfolio are obtained when using the ML methods discussed in Section 3.2 to perform variable selection and regularization. The table also reports the performance of the minimum-variance strategy when using only the market, size, value, profitability, and investment characteristics as well as the performance of the equally weighted and value-weighted strategies. The ML-based portfolios are implemented with two alternative configurations of the data set discussed in Section 2: using only the original 95 predictors used in Gu et al. (2020, 2021) and ii) all 4,610 predictors including the second-order and cross-products transformations. The table also reports the average monthly turnover of each strategy. Panels A and B report results assuming two alternative levels of transaction costs (T.C.): 0 basis points (bp) and 10 bp. One, two, and three asterisks indicate that the differences in portfolio Sharpe ratios with respect to those of the lasso strategy with 4,610 predictors are significant at the 10%, 5% and 1% level, respectively.

	SR before T.C.	Turnover	SR after T.C. (10 bp)
market, size, value	0.690***	0.34	0.662***
market, size, value, profit., invest.	0.991***	0.51	0.949***
equally weighted	0.454	0.16	0.444**
value-weighted	0.509*	0.11	0.500***
OLS	-0.075	5.14	-0.428*
Lasso			
(95 characteristics)	0.112*	1.77	-0.121
(4,610 characteristics)	-0.156	1.14	-0.320
Adalasso			
(95 characteristics)	0.292***	2.31	0.004*
(4,610 characteristics)	0.142**	2.36	-0.169
Elastic net			
(95 characteristics)	0.107*	1.72	-0.119
(4,610 characteristics)	-0.164	1.15	-0.328
Ridge			
(95 characteristics)	0.842***	0.27	0.814***
(4,610 characteristics)	0.557**	0.15	0.545***
Horseshoe			
(95 characteristics)	0.378***	1.07	0.233***
(4,610 characteristics)	0.194**	3.48	-0.251
Scad			
(95 characteristics)	0.087	2.21	-0.178
(4,610 characteristics)	-0.339**	2.46	-0.664***
$L2$-boosting			
(95 characteristics)	1.220***	1.04	1.076***
(4,610 characteristics)	1.153***	1.67	0.925***

Table 4: **Out-of-sample portfolio drawdown and value-at-risk**

The table reports the maximum drawdown and the value-at-risk (VaR) based on the historical simulation method with 99% confidence for the alternative portfolios strategies implemented in the paper.

	Maximun Drawdown	VaR (99%)
market, size, value	52.7%	11.3%
market, size, value, profit., invest.	54.2%	12.2%
equally weighted	56.8%	13.1%
value-weighted	52.9%	11.4%
OLS (95 characteristics)	68.9%	14.8%
Lasso		
(95 characteristics)	53.7%	6.5%
(4,610 characteristics)	55.9%	7.9%
Adalasso		
(95 characteristics)	49.2%	8.3%
(4,610 characteristics)	40.8%	7.7%
Elastic net		
(95 characteristics)	54.1%	7.1%
(4,610 characteristics)	55.3%	7.9%
Ridge		
(95 characteristics)	48.3%	9.8%
(4,610 characteristics)	53.6%	11.4%
Horseshoe		
(95 characteristics)	47.1%	6.7%
(4,610 characteristics)	49.3%	7.4%
Scad		
(95 characteristics)	59.6%	8.2%
(4,610 characteristics)	68.3%	7.1%
$L2$-boosting		
(95 characteristics)	38.6%	7.4%
(4,610 characteristics)	39.2%	6.0%

Table 5: **Predictor selection**

The table presents the average total number of selected predictors for each method, alongside the percentage distribution across predictor classes—main effects, second-order effects, and interaction effects. The "Class-relative percentage" panel shows the proportion of selected predictors within each class. The "Difference from equal-importance" panel displays the deviation of each class's proportion from an equal-importance benchmark, indicating if certain predictor classes are over- or underrepresented relative to an equal distribution. All values are averaged across all estimation rounds.

		Class-relative percentage			Difference from equal-importance		
	Numer of selected predictors	Main effects	Second-order effects	Interaction effects	Main effects	Second-order effects	Interaction effects
Lasso	33	5.5%	7.2%	87.3%	3.5%	5.2%	-8.7%
Adalasso	62	0.8%	2.5%	96.7%	-1.3%	0.6%	0.7%
Elastic net	60	4.8%	6.5%	88.7%	2.7%	4.6%	-7.3%
Horseshoe	47	3.5%	2.8%	93.7%	1.5%	0.9%	-2.3%
Scad	16	16.0%	4.4%	79.5%	14.0%	2.5%	-16.5%
L2-boosting	75	1.4%	1.8%	96.8%	-0.7%	-0.2%	0.8%

Table 6: **Economic significance of the most important predictors**

This table reports the regression results of selected predictor coefficients on state variables, including the investor sentiment index from Baker and Wurgler (2006), the liquidity index from Pástor and Stambaugh (2003), and the NBER recession indicator. The results are presented for six different estimation methods: Lasso, Adalasso, Elastic-Net, Scad, L2-Boosting, and Horseshoe. Panel A reports the results for the main effects of each predictor, while Panel B shows the interaction effects. One, two, and three asterisks indicate that the estimated coefficient is significant at the 10%, 5% and 1% level, respectively. A dash ("-") indicates that the predictor was not selected by the method or that the estimation coefficient is equal to zero.

	Lasso	Adalasso	Elastic-net	Scad	L2-Boosting	Horseshoe
Panel A: Main effects						
beta						
sentiment	-0.208***	-0.097***	-0.152***	-0.475***	-0.135***	-0.312**
recession	-0.087	0.025	0.155	0.077	-0.173***	0.205
liquidity	-0.127	-0.058	-0.060	0.222	-0.199	0.166
dolvol						
sentiment	-	-	0.001	0.031	-	-0.004
recession	-	-	0.002	0.064	-	0.003
liquidity	-0.001	-	0.001	-0.016	-	-0.074
idiovol						
sentiment	0.022	-	0.003	-0.067	-	0.068*
recession	-0.040	-	-0.045*	0.047	-	0.051
liquidity	0.008	-	0.004	-0.264	-	0.205
Panel B: Interactions						
beta__x__dolvol						
sentiment	-0.012***	0.020***	-0.014*	-0.038***	-	0.055
recession	-0.017***	0.019	-0.012	-0.059**	-	0.086
liquidity	-0.022	-0.083	-0.03	-0.170	-	-0.082
dolvol__x__idiovol						
sentiment	0.184***	0.001	0.198***	0.219***	0.046***	0.027
recession	0.005	0.001	-0.016	0.016	0.023	0.076
liquidity	0.257	-0.004	0.154	0.072	-0.115	-0.477
beta__x__idiovol						
sentiment	-	-	-0.007	-0.013	0.050***	0.024
recession	-	-	-0.048***	-0.011	0.052***	0.008
liquidity	-	-	-0.040	-0.134	-0.243**	-0.050

Table 7: **Portfolio weight statistics**

The table reports descriptive statistics of the distribution of portfolio weights of the various strategies considered. The columns reports the time series averages of the minimum weight, maximum weight, and fraction of negative weights (short proportion).

	Minimum weight (%)	Maximum weight (%)	Short proportion (%)
market, size, value	-0.143	0.578	0.426
market, size, value, profit., invest.	-0.352	0.543	0.379
equally weighted	0.000	0.067	0.000
value-weighted	0.000	0.432	0.000
OLS (95 characteristics)	-1.760	1.753	0.442
Lasso			
(95 characteristics)	-0.422	0.477	0.379
(4,610 characteristics)	-0.318	0.395	0.329
Adalasso			
(95 characteristics)	-0.663	0.710	0.386
(4,610 characteristics)	-0.834	0.821	0.333
Elastic net			
(95 characteristics)	-0.401	0.443	0.371
(4,610 characteristics)	-0.315	0.382	0.324
Ridge			
(95 characteristics)	-0.105	0.075	0.152
(4,610 characteristics)	-0.029	0.051	0.022
Horseshoe			
(95 characteristics)	-0.297	0.532	0.384
(4,610 characteristics)	-0.494	0.825	0.431
Scad			
(95 characteristics)	-0.431	0.566	0.397
(4,610 characteristics)	-0.446	0.534	0.371
$L2$-boosting			
(95 characteristics)	-0.406	0.454	0.325
(4,610 characteristics)	-0.696	0.657	0.328

Table 8: **Performance of minimum-variance parametric portfolios with marginal screening**

The table reports annualized out-of-sample (OOS) standard deviation of the minimum-variance parametric portfolio returns when the marginal screening (MS) method of Fan and Lv (2008) is used to screen the pool of 4,610 predictors to select the best 5, 10, and 50 predictors. The table also reports the average monthly turnover of each strategy. Panels A and B report results assuming two alternative levels of transaction costs: 0 basis points (b.p.) and 10 b.p. One, two, and three asterisks indicate that the differences in portfolio variance and in Sharpe ratios with respect to those of the lasso strategy with 4,610 predictors are significant at the 10%, 5% and 1% level, respectively.

	Std. dev. of OOS portfolio returns
Top-3 predictors	0.107***
Top-5 predictors	0.105***
Top-10 predictors	0.095**
Top-50 predictors	0.103***

Figure 1: **Predictor returns**

The figure plots the average values of the aggregate monthly returns for the predictors selected with each method. Average values of $\theta^{\intercal} r_c$ calculated across all rolling estimation windows.

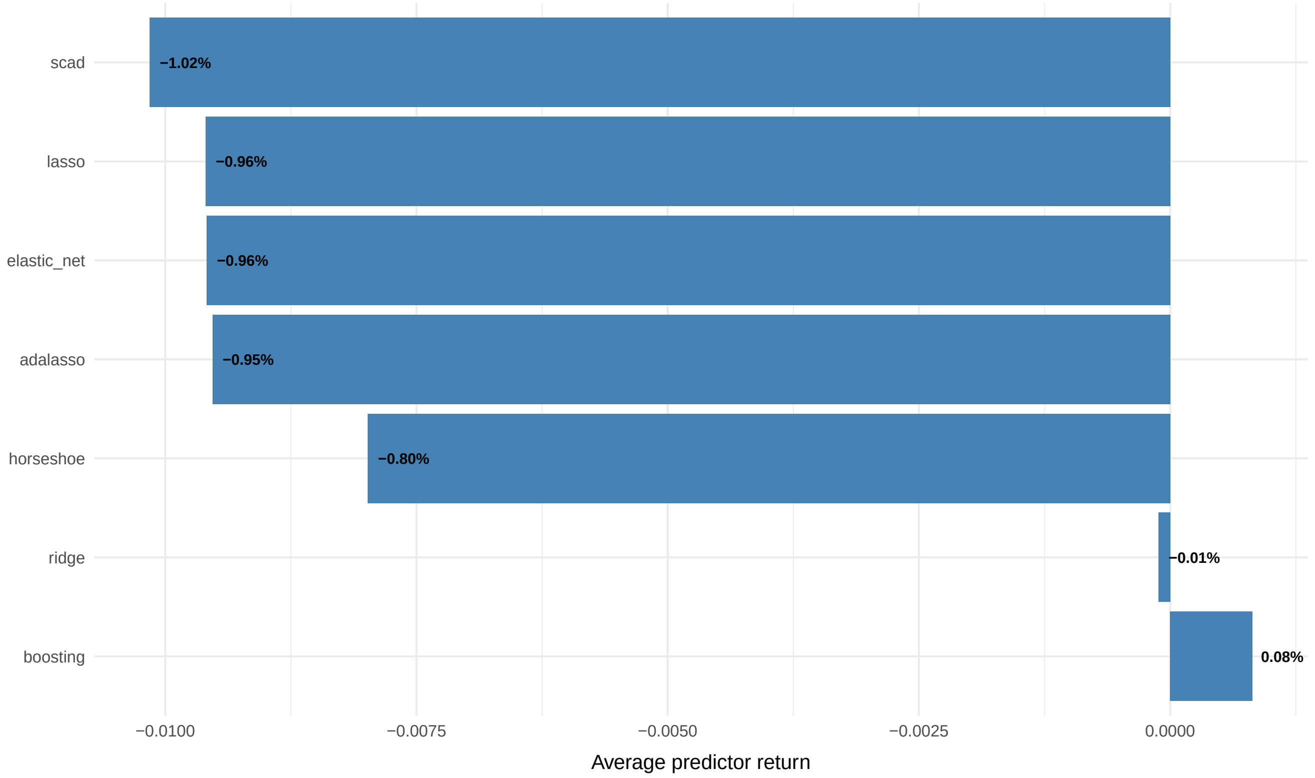


Figure 2: **Individual predictor returns**

The figure plots the returns on the top-10 and bottom-10 individual predictors for each method.

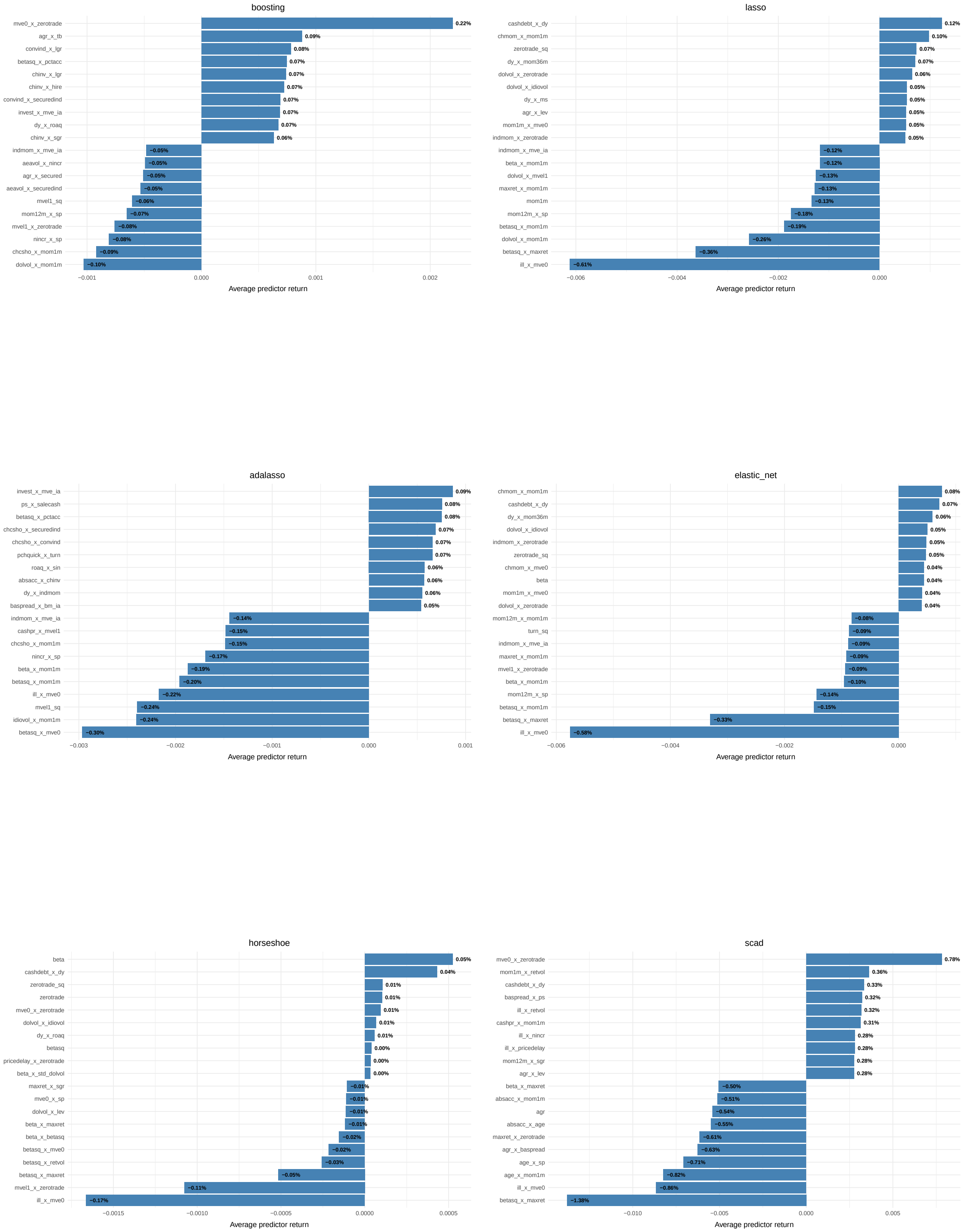


Figure 3: **Lasso predictor importance**

The figure plots marginal contributions of the top-20 selected predictors for the lasso method. The figure breaks down the contributions into three components: the predictor's own variance, its covariance with other predictors, and its correlation with the benchmark portfolios (see Eq. (17)). Importance based on own-variance (left-hand plot) is ranked in ascending order: predictors whose own variance leads to smaller increases in the objective function (Eq. (16)) are more important. Importance based on covariance with other predictors (center plot) and correlation with the benchmark portfolio (right-hand plot) is ranked in descending order: predictors with lower covariance with others and with the benchmark portfolio contribute more to decrease the objective function and are thus more important. Finally, predictors with a positive (negative) parametric-portfolio coefficient are in blue (red).

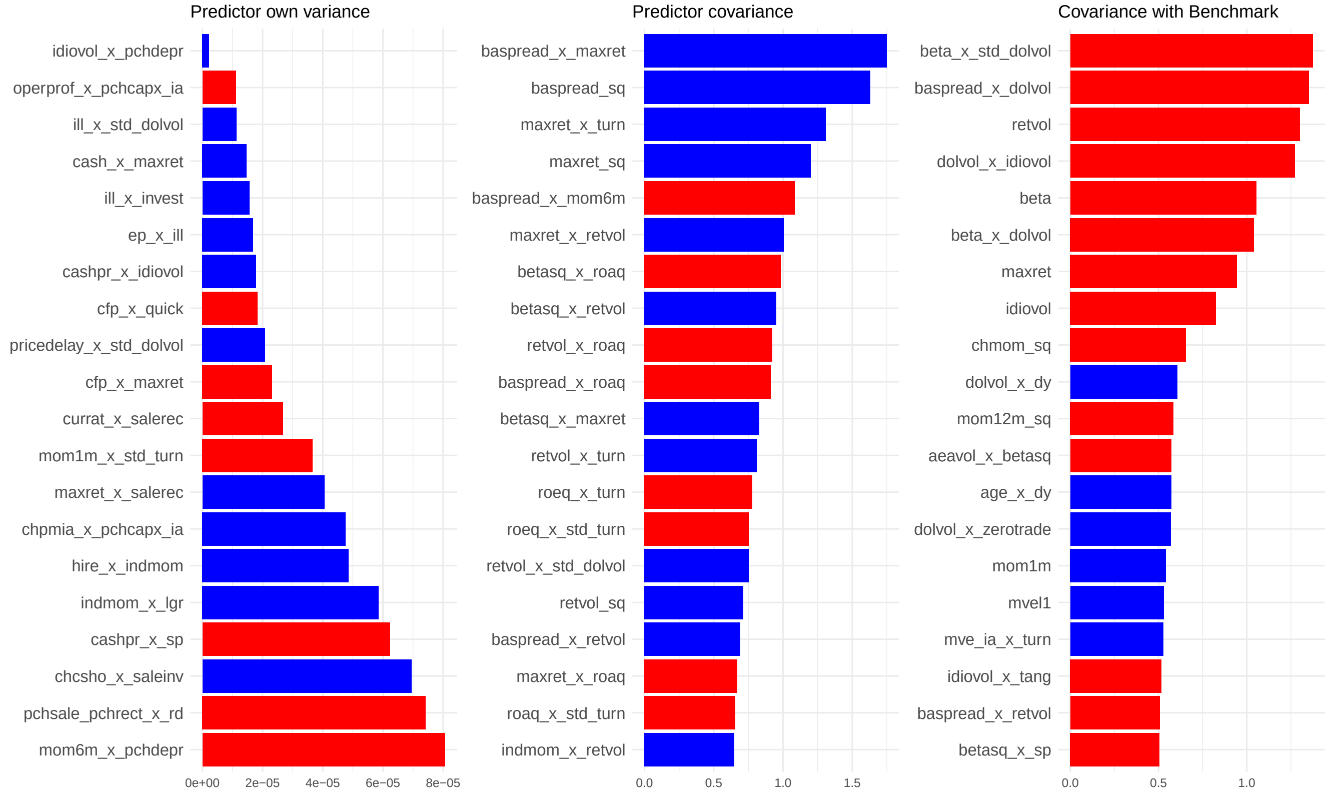


Figure 4: **Elastic net predictor importance**

The figure plots marginal contributions of the top-20 selected predictors for the elastic net method. The figure breaks down the contributions into three components: the predictor's own variance, its covariance with other predictors, and its correlation with the benchmark portfolios (see Eq. (17)). Importance based on own-variance (left-hand plot) is ranked in ascending order: predictors whose own variance leads to smaller increases in the objective function (Eq. (16)) are more important. Importance based on covariance with other predictors (center plot) and correlation with the benchmark portfolio (right-hand plot) is ranked in descending order: predictors with lower covariance with others and with the benchmark portfolio contribute more to decrease the objective function and are thus more important. Finally, predictors with a positive (negative) parametric-portfolio coefficient are in blue (red).

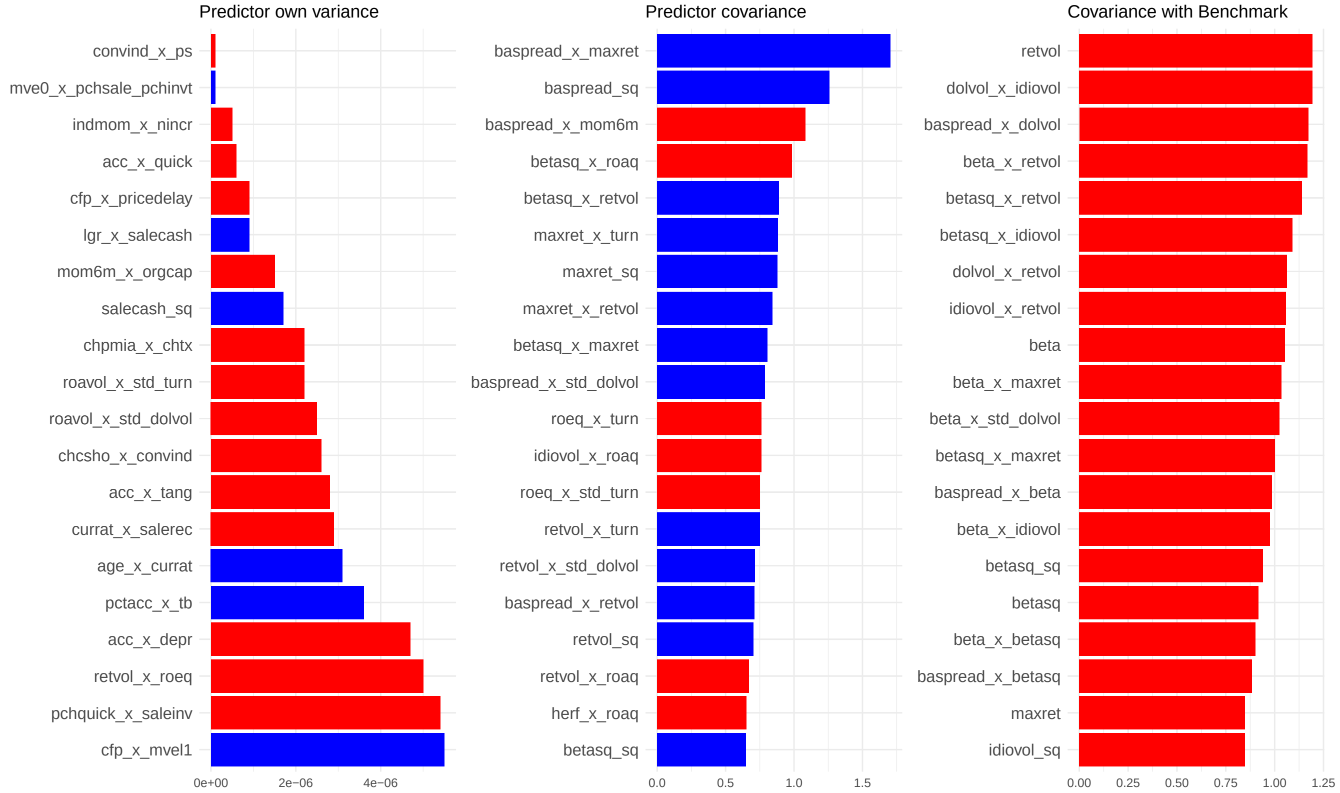


Figure 5: $L2$-boosting predictor importance

The figure plots marginal contributions of the top-20 selected predictors for the $L2$-boosting method. The figure breaks down the contributions into three components: the predictor's own variance, its covariance with other predictors, and its correlation with the benchmark portfolios (see Eq. (17)). Importance based on own-variance (left-hand plot) is ranked in ascending order: predictors whose own variance leads to smaller increases in the objective function (Eq. (16)) are more important. Importance based on covariance with other predictors (center plot) and correlation with the benchmark portfolio (right-hand plot) is ranked in descending order: predictors with lower covariance with others and with the benchmark portfolio contribute more to decrease the objective function and are thus more important. Finally, predictors with a positive (negative) parametric-portfolio coefficient are in blue (red).

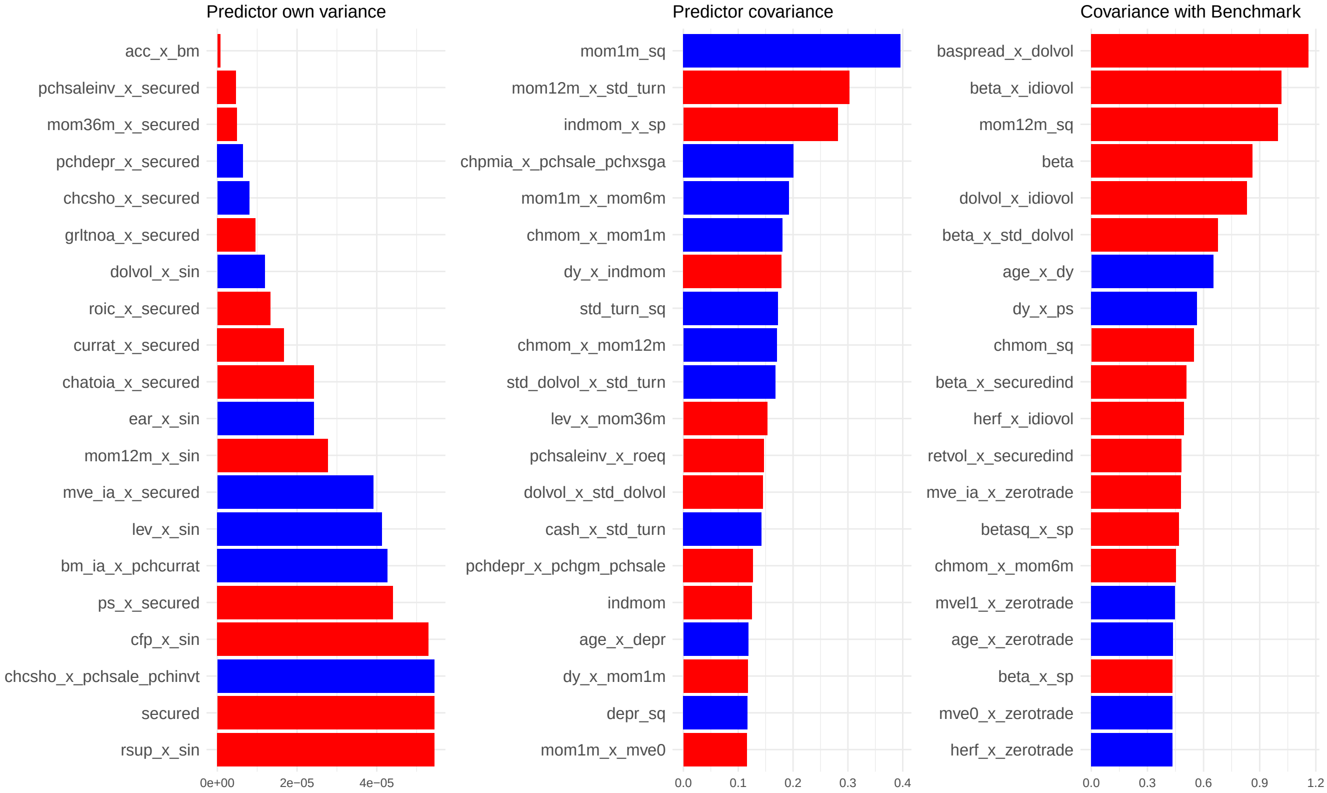
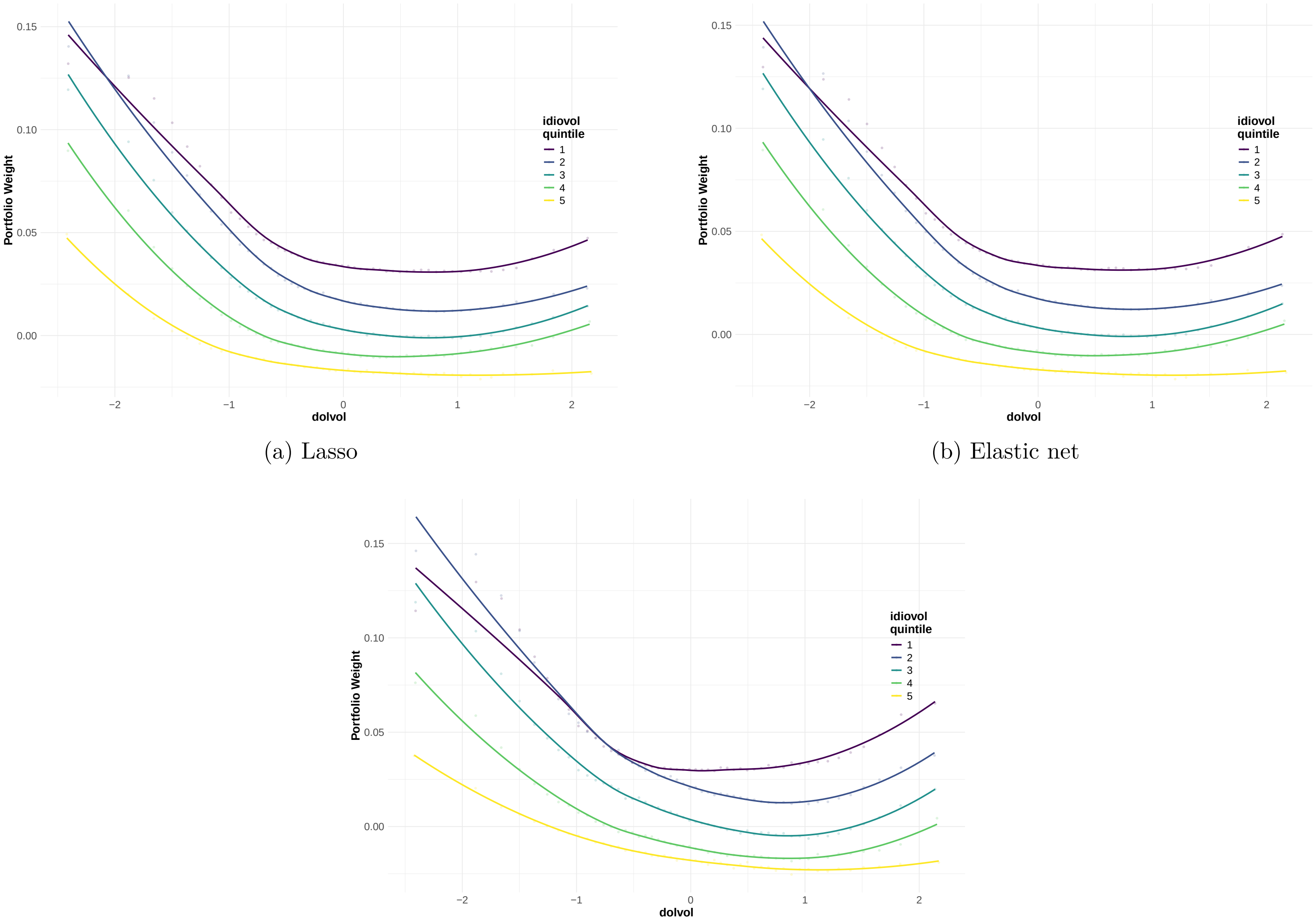


Figure 6: **Portfolio weights, liquidity and idiosyncratic volatility**

The figure plots the average minimum-variance portfolio weights (vertical axis) across values of the standardized liquidity (*dolvol*) and across quintiles of idiosyncratic volatility (*idiovol*).



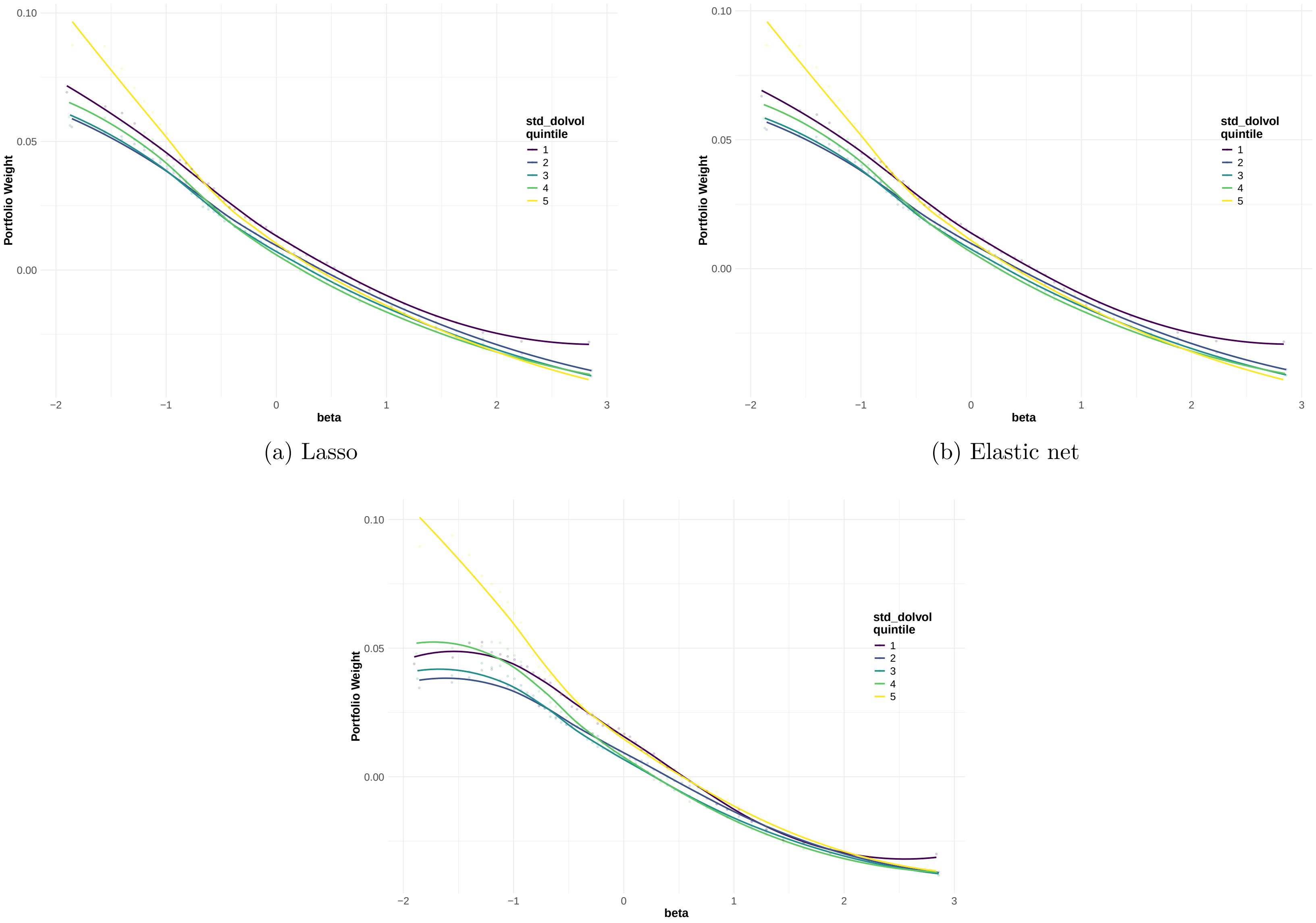
(a) Lasso

(b) Elastic net

(c) $L2$-boosting

Figure 7: **Portfolio weights, market beta and standard deviation of liquidity**

The figure plots the average minimum-variance portfolio weights (vertical axis) across values of the standardized market beta (*beta*) and across quintiles of standard deviation of liquidity (*std_dolvol*).



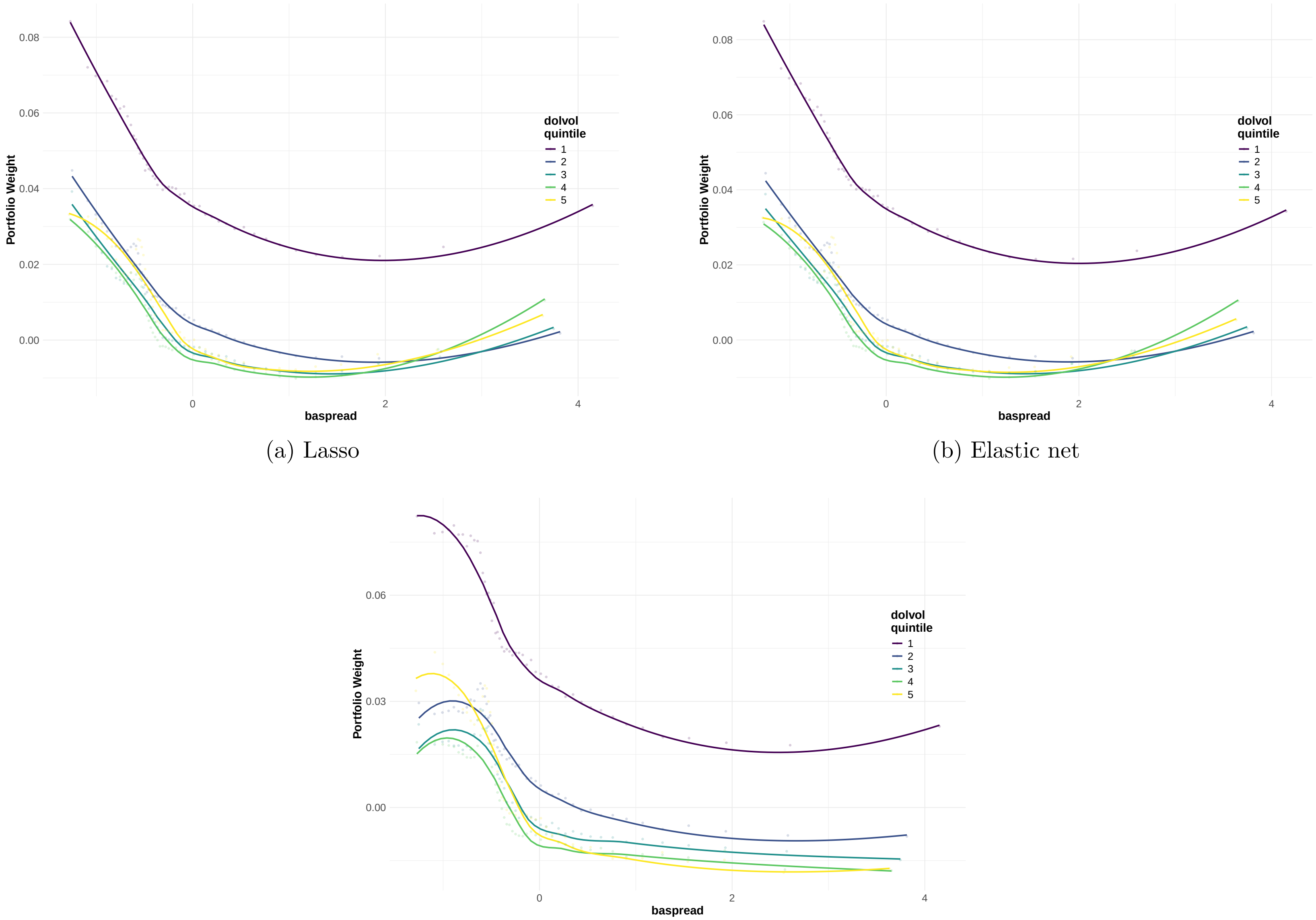
(a) Lasso

(b) Elastic net

(c) *L*2-boosting

Figure 8: **Portfolio weights, bid-ask spread and liquidity**

The figure plots the average minimum-variance portfolio weights (vertical axis) across values of the standardized bid-ask spread (*baspread*) and across quintiles of liquidity (*dolvol*).



(a) Lasso

(b) Elastic net

(c) *L*2-boosting

Figure 9: **Lasso coefficients: first-order effects**

The figure plots the estimated coefficients of the features' first-order effects when using the lasso method across all estimation rounds.

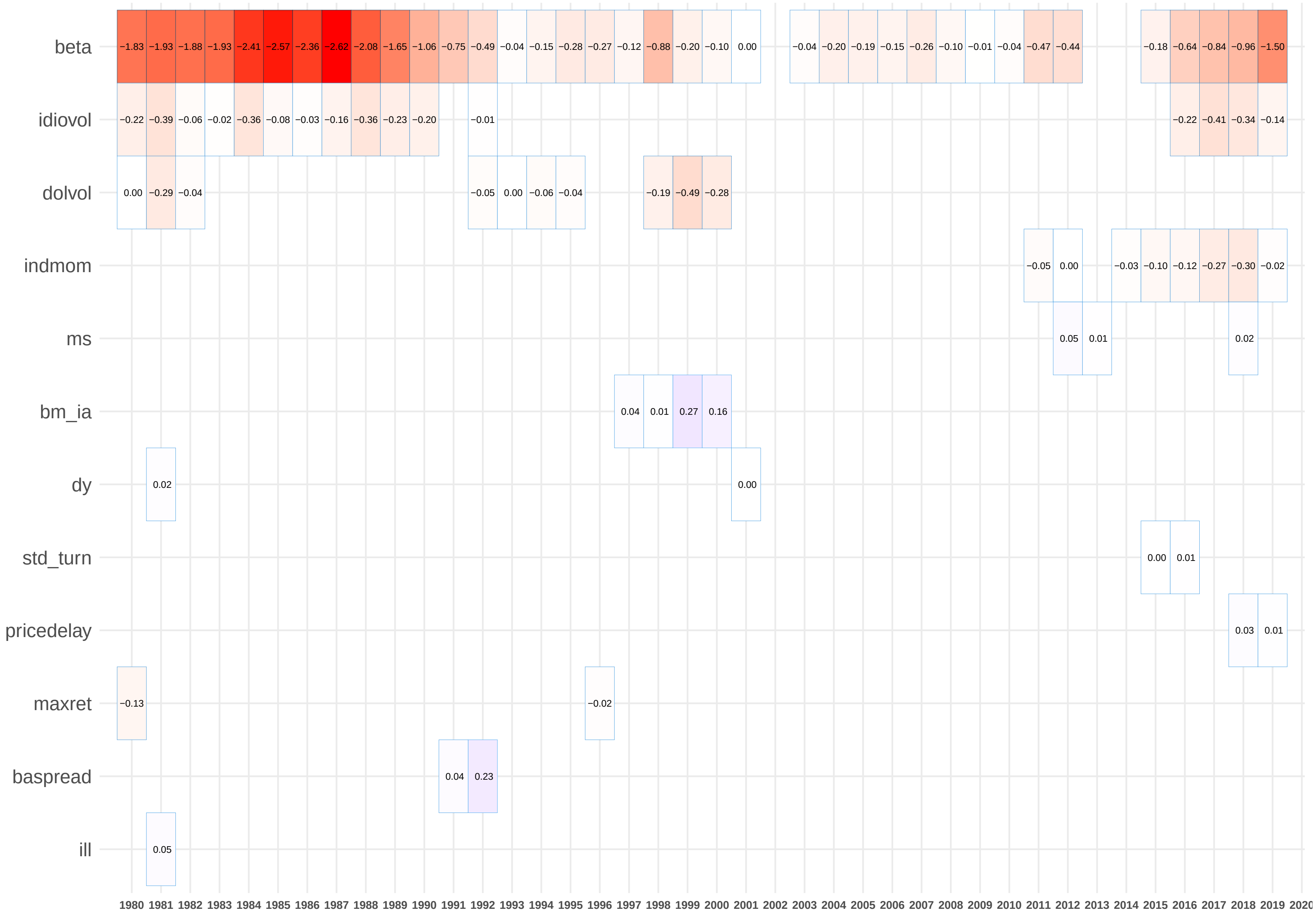


Figure 10: **Lasso coefficients: second-order effects**

The figure plots the estimated coefficients of the features' second-order effects when using the lasso method across all estimation rounds.

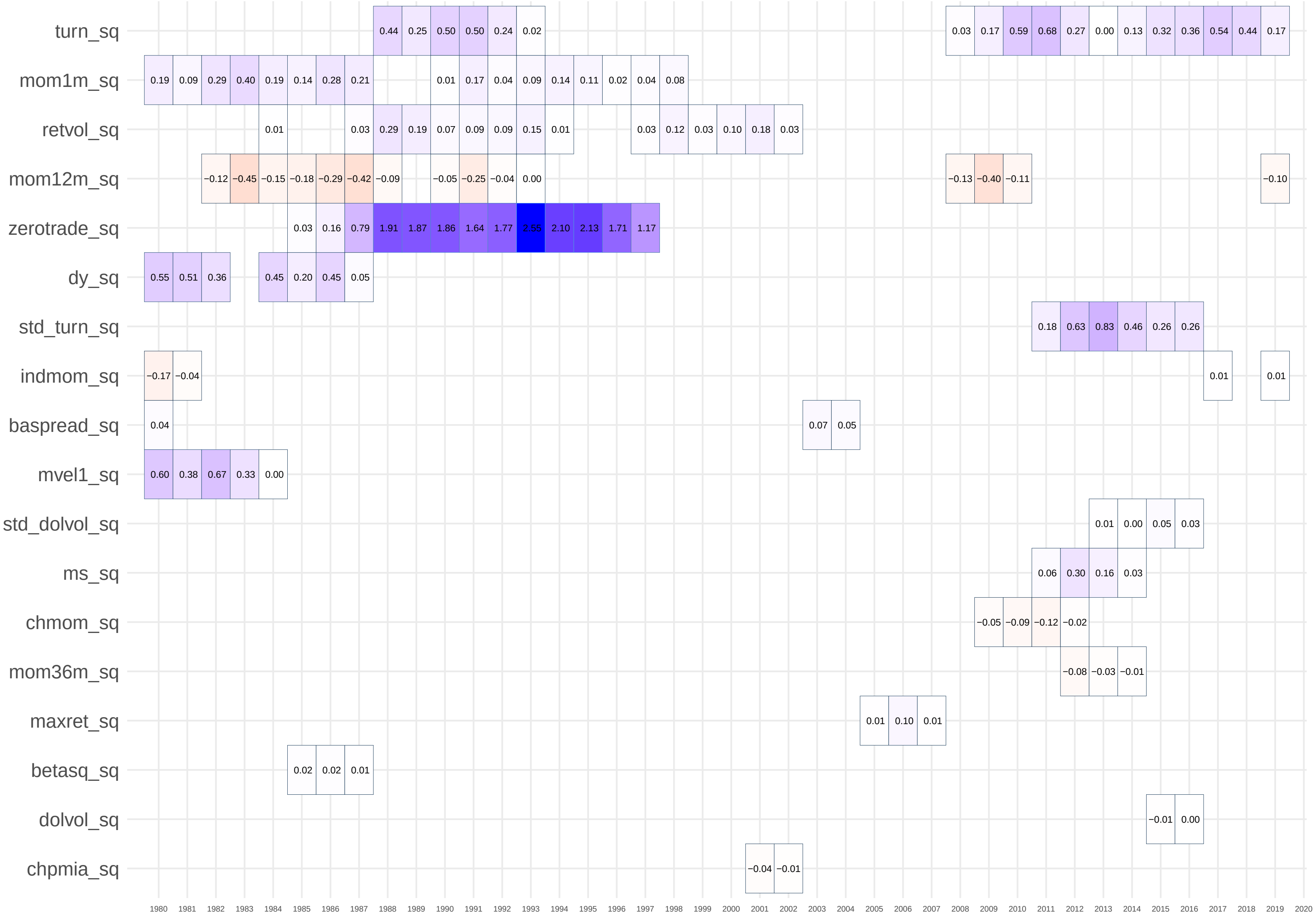


Figure 11: **Lasso coefficients: interactions**

The figure plots the estimated coefficients of the features' interaction effects when using the lasso method across all estimation rounds.

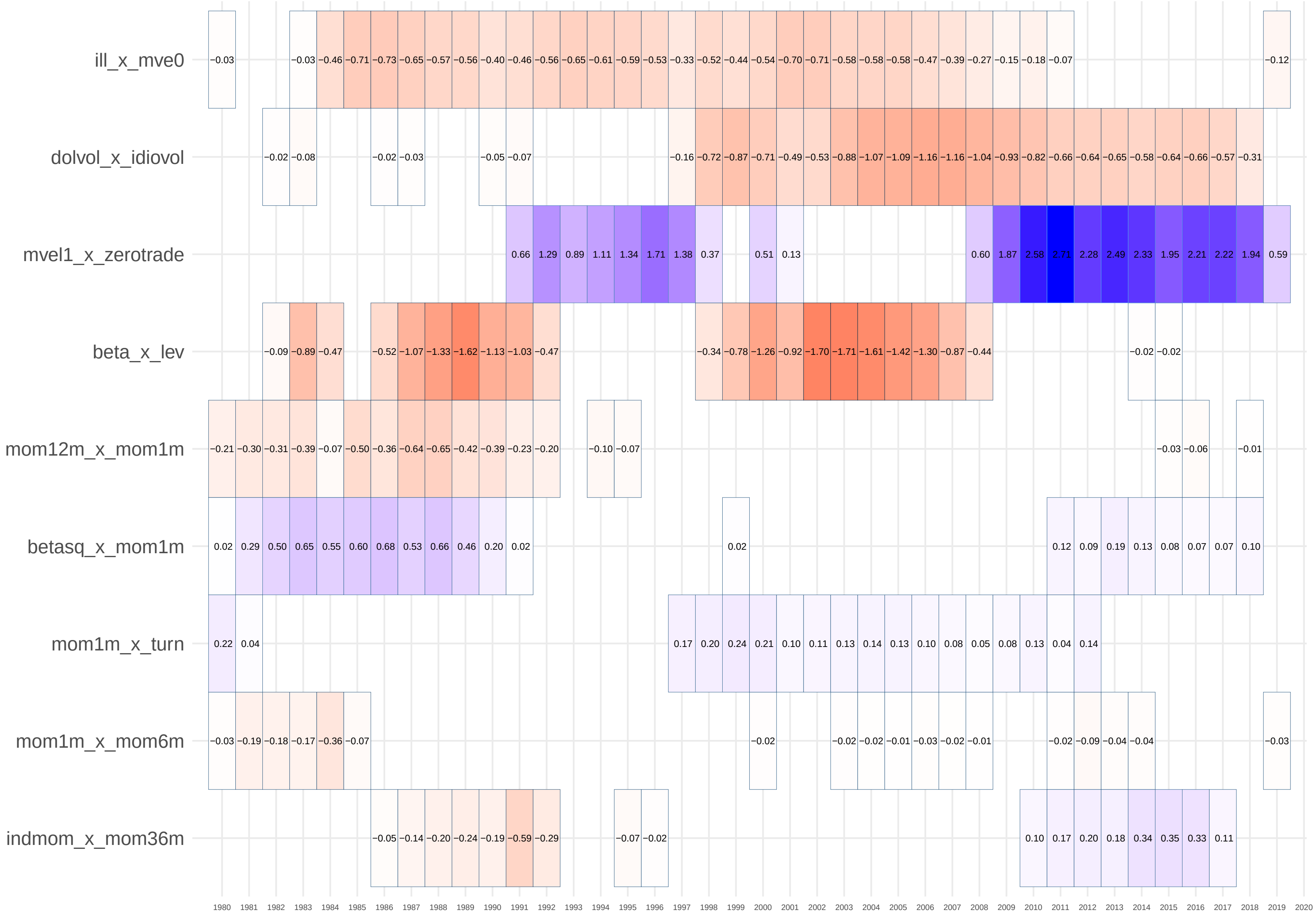
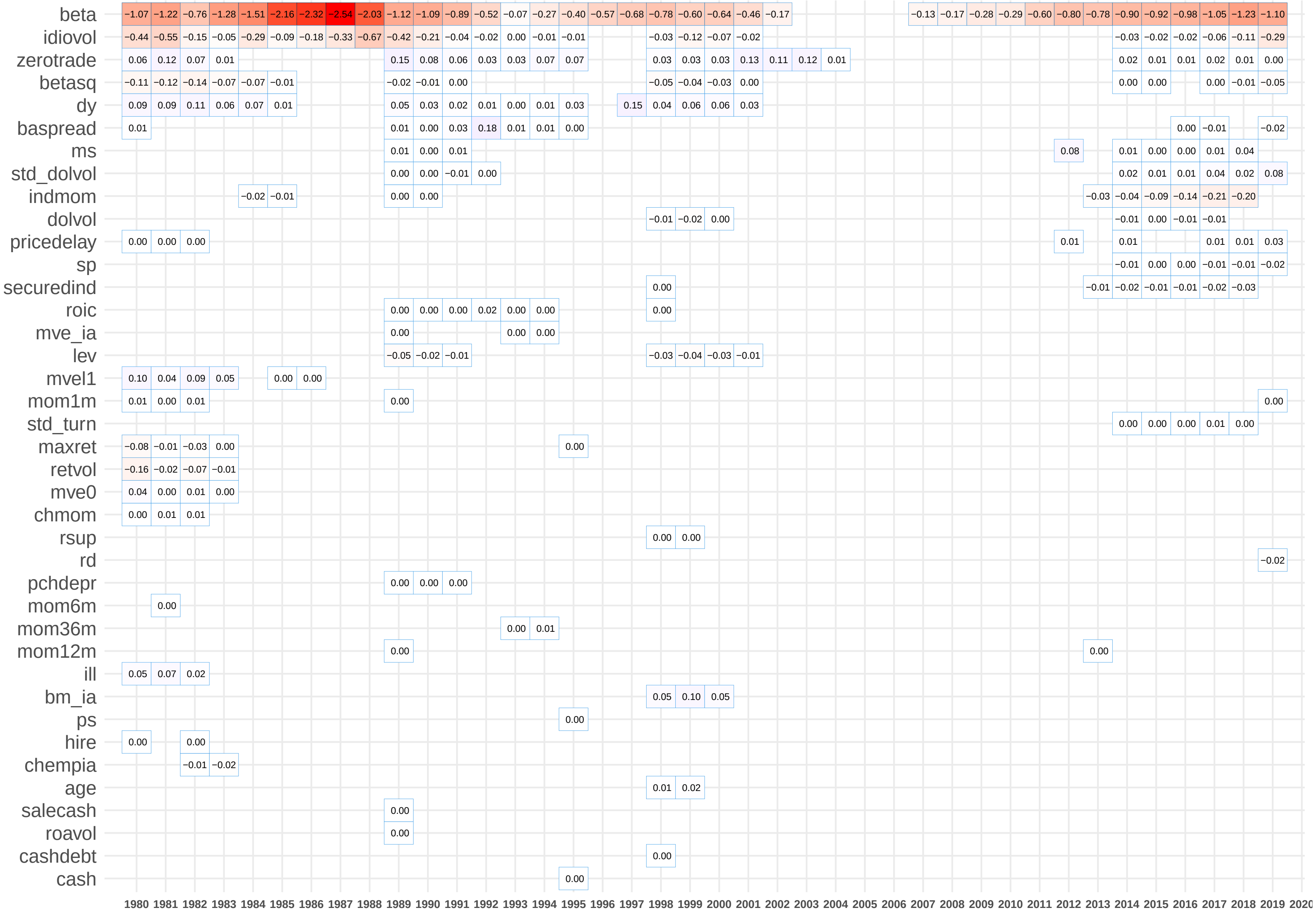


Figure 12: **Elastic net coefficients: first-order effects**

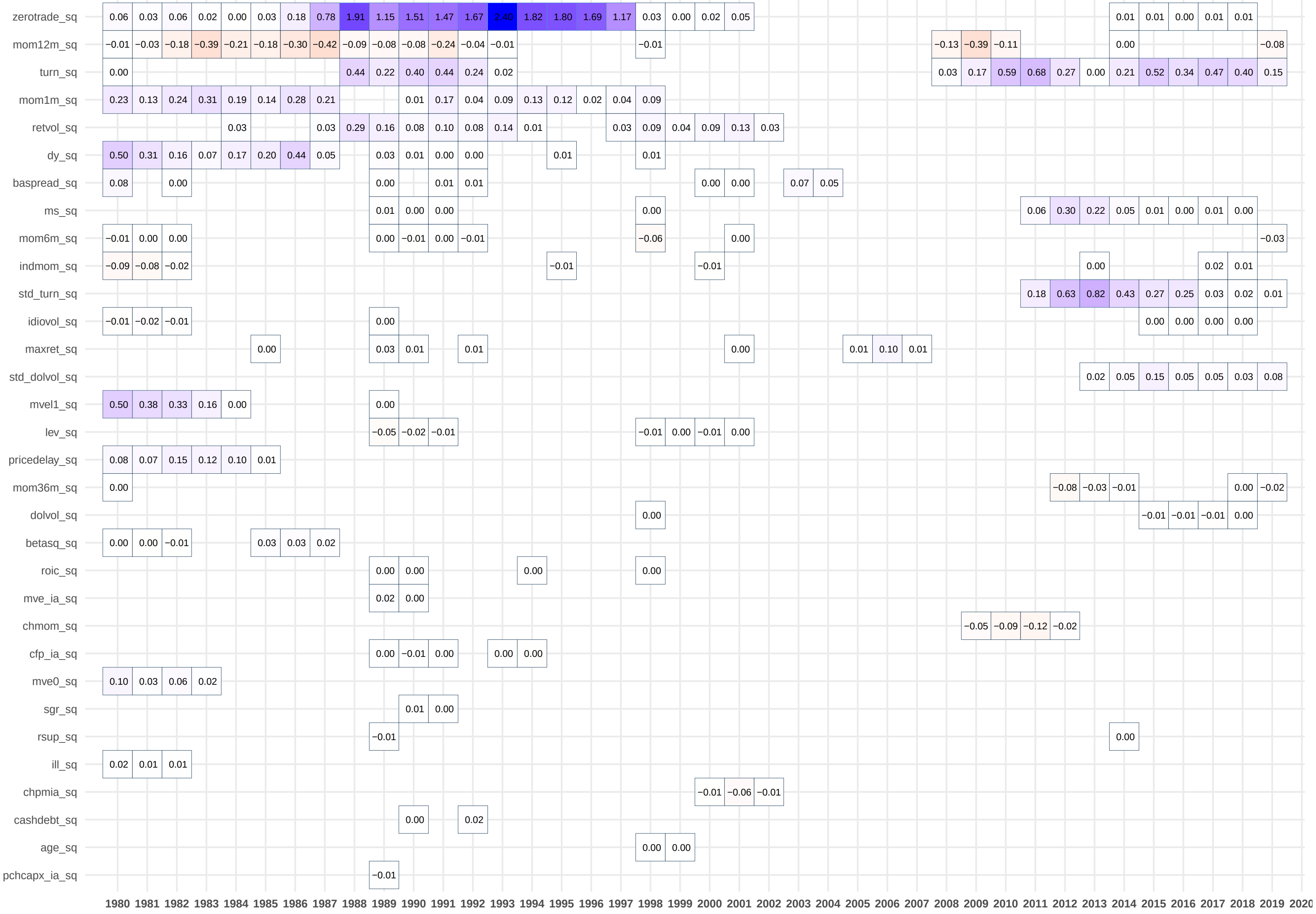
The figure plots the estimated coefficients of the feature's first-order effects when using the elastic net method across all estimation rounds.



	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
beta	−1.07	−1.22	−0.76	−1.28	−1.51	−2.16	−2.32	−2.54	−2.03	−1.12	−1.09	−0.89	−0.52	−0.07	−0.27	−0.40	−0.57	−0.68	−0.78	−0.60	−0.64	−0.46	−0.17					−0.13	−0.17	−0.28	−0.29	−0.60	−0.80	−0.78	−0.90	−0.92	−0.98	−1.05	−1.23	−1.10	
idiovol	−0.44	−0.55	−0.15	−0.05	−0.29	−0.09	−0.18	−0.33	−0.67	−0.42	−0.21	−0.04	−0.02	0.00	−0.01	−0.01			−0.03	−0.12	−0.07	−0.02													−0.03	−0.02	−0.02	−0.06	−0.11	−0.29	
zerotrade	0.06	0.12	0.07	0.01						0.15	0.08	0.06	0.03	0.03	0.07	0.07			0.03	0.03	0.03	0.13	0.11	0.12	0.01										0.02	0.01	0.01	0.02	0.01	0.00	
betasq	−0.11	−0.12	−0.14	−0.07	−0.07	−0.01				−0.02	−0.01	0.00							−0.05	−0.04	−0.03	0.00													0.00	0.00		0.00	−0.01	−0.05	
dy	0.09	0.09	0.11	0.06	0.07	0.01				0.05	0.03	0.02	0.01	0.00	0.01	0.03		0.15	0.04	0.06	0.06	0.03																			
baspread	0.01									0.01	0.00	0.03	0.18	0.01	0.01	0.00																					0.00	−0.01		−0.02	
ms										0.01	0.00	0.01																					0.08		0.01	0.00	0.00	0.01	0.04		
std_dolvol										0.00	0.00	−0.01	0.00																						0.02	0.01	0.01	0.04	0.02	0.08	
indmom					−0.02	−0.01				0.00	0.00																							−0.03	−0.04	−0.09	−0.14	−0.21	−0.20		
dolvol																			−0.01	−0.02	0.00														−0.01	0.00	−0.01	−0.01			
pricedelay	0.00	0.00	0.00																														0.01		0.01			0.01	0.01	0.03	
sp																																			−0.01	0.00	0.00	−0.01	−0.01	−0.02	
securedind																			0.00															−0.01	−0.02	−0.01	−0.01	−0.02	−0.03		
roic										0.00	0.00	0.00	0.02	0.00	0.00				0.00																						
mve_ia										0.00				0.00	0.00																										
lev										−0.05	−0.02	−0.01							−0.03	−0.04	−0.03	−0.01																			
mvel1	0.10	0.04	0.09	0.05		0.00	0.00																																		
mom1m	0.01	0.00	0.01							0.00																														0.00	
std_turn																																			0.00	0.00	0.00	0.01	0.00		
maxret	−0.08	−0.01	−0.03	0.00												0.00																									
retvol	−0.16	−0.02	−0.07	−0.01																																					
mve0	0.04	0.00	0.01	0.00																																					
chmom	0.00	0.01	0.01																																						
rsup																			0.00	0.00																					
rd																																								−0.02	
pchdepr										0.00	0.00	0.00																													
mom6m		0.00																																							
mom36m														0.00	0.01																										
mom12m										0.00																								0.00							
ill	0.05	0.07	0.02																																						
bm_ia																			0.05	0.10	0.05																				
ps																0.00																									
hire	0.00		0.00																																						
chempia			−0.01	−0.02																																					
age																			0.01	0.02																					
salecash										0.00																															
roavol										0.00																															
cashdebt																			0.00																						
cash																0.00																									

Figure 13: **Elastic net coefficients: second-order effects**

The figure plots the estimated coefficients of the feature's second-order effects when using the elastic net method across all estimation rounds.

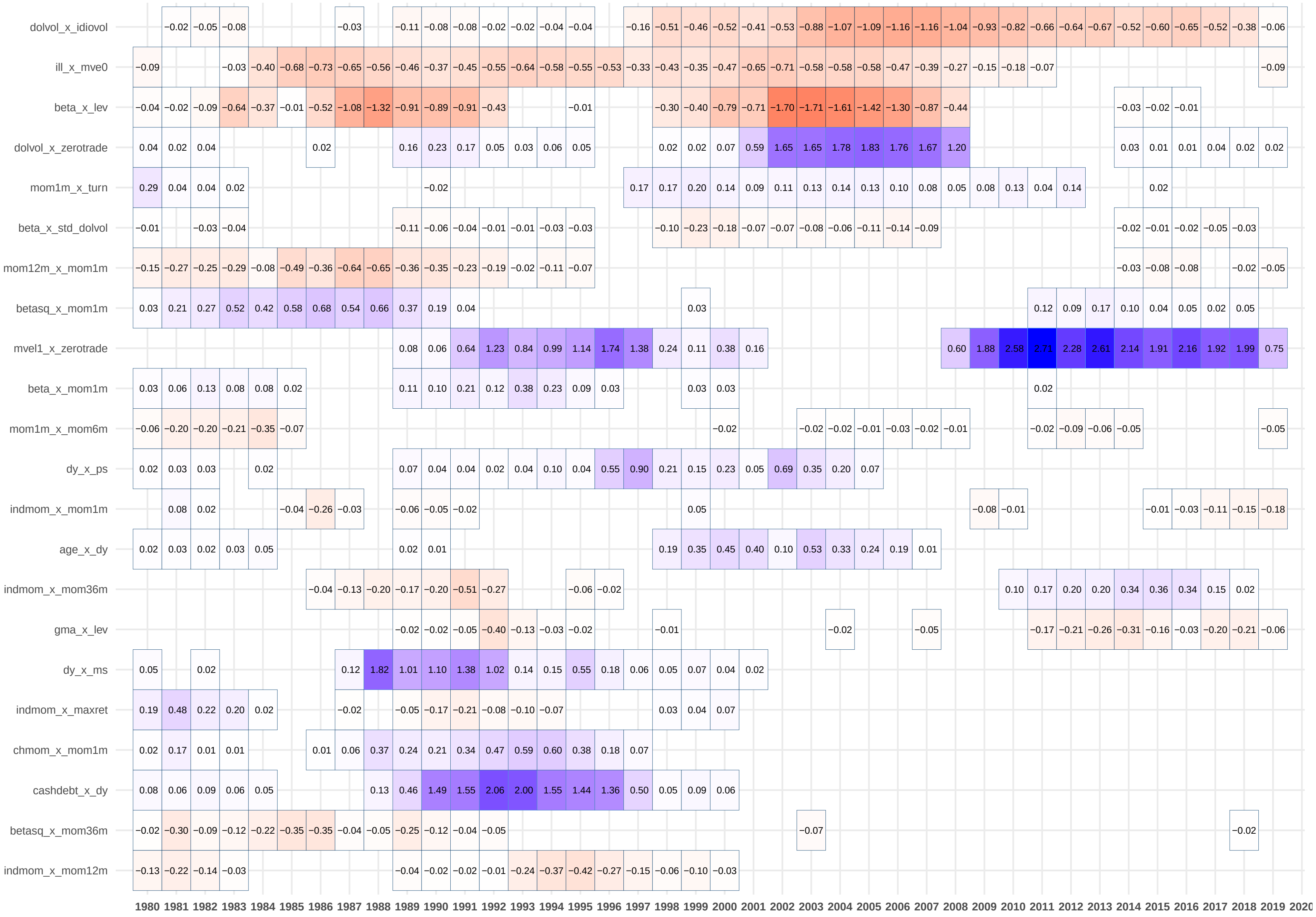


	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
zerotrade_sq	0.06	0.03	0.06	0.02	0.00	0.03	0.18	0.78	1.91	1.15	1.51	1.47	1.67	2.40	1.82	1.80	1.69	1.17	0.03	0.00	0.02
mom12m_sq	−0.01	−0.03	−0.18	−0.39	−0.21	−0.18	−0.30	−0.42	−0.09	−0.08	−0.08	−0.24	−0.04	−0.01					−0.01		
turn_sq	0.00								0.44	0.22	0.40	0.44	0.24	0.02							
mom1m_sq	0.23	0.13	0.24	0.31	0.19	0.14	0.28	0.21			0.01	0.17	0.04	0.09	0.13	0.12	0.02	0.04	0.09		
retvol_sq					0.03			0.03	0.29	0.16	0.08	0.10	0.08	0.14	0.01			0.03	0.09	0.04	0.09
dy_sq	0.50	0.31	0.16	0.07	0.17	0.20	0.44	0.05		0.03	0.01	0.00	0.00			0.01			0.01		
baspread_sq	0.08		0.00							0.00		0.01	0.01								0.00
ms_sq										0.01	0.00	0.00							0.00		
mom6m_sq	−0.01	0.00	0.00							0.00	−0.01	0.00	−0.01						−0.06		
indmom_sq	−0.09	−0.08	−0.02													−0.01					−0.01
std_turn_sq																					
idiovol_sq	−0.01	−0.02	−0.01							0.00											
maxret_sq						0.00				0.03	0.01		0.01								
std_dolvol_sq																					
mvel1_sq	0.50	0.38	0.33	0.16	0.00					0.00											
lev_sq										−0.05	−0.02	−0.01							−0.01	0.00	−0.01
pricedelay_sq	0.08	0.07	0.15	0.12	0.10	0.01															
mom36m_sq	0.00																				
dolvol_sq																			0.00		
betasq_sq	0.00	0.00	−0.01			0.03	0.03	0.02													
roic_sq										0.00	0.00				0.00				0.00		
mve_ia_sq										0.02	0.00										
chmom_sq																					
cfp_ia_sq										0.00	−0.01	0.00		0.00	0.00						
mve0_sq	0.10	0.03	0.06	0.02																	
sgr_sq											0.01	0.00									
rsup_sq										−0.01											
ill_sq	0.02	0.01	0.01																		
chpmia_sq																					−0.01
cashdebt_sq											0.00		0.02								
age_sq																			0.00	0.00	
pchcapx_ia_sq										−0.01											

	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
zerotrade_sq	0.05													0.01	0.01	0.00	0.01	0.01		
mom12m_sq								−0.13	−0.39	−0.11				0.00					−0.08	
turn_sq								0.03	0.17	0.59	0.68	0.27	0.00	0.21	0.52	0.34	0.47	0.40	0.15	
mom1m_sq																				
retvol_sq	0.13	0.03																		
dy_sq																				
baspread_sq	0.00		0.07	0.05																
ms_sq											0.06	0.30	0.22	0.05	0.01	0.00	0.01	0.00		
mom6m_sq	0.00																		−0.03	
indmom_sq													0.00				0.02	0.01		
std_turn_sq											0.18	0.63	0.82	0.43	0.27	0.25	0.03	0.02	0.01	
idiovol_sq															0.00	0.00	0.00	0.00		
maxret_sq	0.00				0.01	0.10	0.01													
std_dolvol_sq													0.02	0.05	0.15	0.05	0.05	0.03	0.08	
mvel1_sq																				
lev_sq	0.00																			
pricedelay_sq																				
mom36m_sq												−0.08	−0.03	−0.01				0.00	−0.02	
dolvol_sq															−0.01	−0.01	−0.01	0.00		
betasq_sq																				
roic_sq																				
mve_ia_sq																				
chmom_sq									−0.05	−0.09	−0.12	−0.02								
cfp_ia_sq																				
mve0_sq																				
sgr_sq																				
rsup_sq														0.00						
ill_sq																				
chpmia_sq	−0.06	−0.01																		
cashdebt_sq																				
age_sq																				
pchcapx_ia_sq																				

Figure 14: **Elastic net coefficients: interactions**

The figure plots the estimated coefficients of the feature's interactions when using the elastic net method across all estimation rounds.



	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
dolvol_x_idiovol		−0.02	−0.05	−0.08				−0.03		−0.11	−0.08	−0.08	−0.02	−0.02	−0.04	−0.04		−0.16	−0.51	−0.46
ill_x_mve0	−0.09			−0.03	−0.40	−0.68	−0.73	−0.65	−0.56	−0.46	−0.37	−0.45	−0.55	−0.64	−0.58	−0.55	−0.53	−0.33	−0.43	−0.35
beta_x_lev	−0.04	−0.02	−0.09	−0.64	−0.37	−0.01	−0.52	−1.08	−1.32	−0.91	−0.89	−0.91	−0.43			−0.01			−0.30	−0.40
dolvol_x_zerotrade	0.04	0.02	0.04				0.02			0.16	0.23	0.17	0.05	0.03	0.06	0.05			0.02	0.02
mom1m_x_turn	0.29	0.04	0.04	0.02							−0.02							0.17	0.17	0.20
beta_x_std_dolvol	−0.01		−0.03	−0.04						−0.11	−0.06	−0.04	−0.01	−0.01	−0.03	−0.03			−0.10	−0.23
mom12m_x_mom1m	−0.15	−0.27	−0.25	−0.29	−0.08	−0.49	−0.36	−0.64	−0.65	−0.36	−0.35	−0.23	−0.19	−0.02	−0.11	−0.07				
betasq_x_mom1m	0.03	0.21	0.27	0.52	0.42	0.58	0.68	0.54	0.66	0.37	0.19	0.04								0.03
mvel1_x_zerotrade										0.08	0.06	0.64	1.23	0.84	0.99	1.14	1.74	1.38	0.24	0.11
beta_x_mom1m	0.03	0.06	0.13	0.08	0.08	0.02				0.11	0.10	0.21	0.12	0.38	0.23	0.09	0.03			0.03
mom1m_x_mom6m	−0.06	−0.20	−0.20	−0.21	−0.35	−0.07														
dy_x_ps	0.02	0.03	0.03		0.02					0.07	0.04	0.04	0.02	0.04	0.10	0.04	0.55	0.90	0.21	0.15
indmom_x_mom1m		0.08	0.02			−0.04	−0.26	−0.03		−0.06	−0.05	−0.02								0.05
age_x_dy	0.02	0.03	0.02	0.03	0.05					0.02	0.01								0.19	0.35
indmom_x_mom36m							−0.04	−0.13	−0.20	−0.17	−0.20	−0.51	−0.27			−0.06	−0.02			
gma_x_lev										−0.02	−0.02	−0.05	−0.40	−0.13	−0.03	−0.02			−0.01	
dy_x_ms	0.05		0.02					0.12	1.82	1.01	1.10	1.38	1.02	0.14	0.15	0.55	0.18	0.06	0.05	0.07
indmom_x_maxret	0.19	0.48	0.22	0.20	0.02			−0.02		−0.05	−0.17	−0.21	−0.08	−0.10	−0.07				0.03	0.04
chmom_x_mom1m	0.02	0.17	0.01	0.01			0.01	0.06	0.37	0.24	0.21	0.34	0.47	0.59	0.60	0.38	0.18	0.07		
cashdebt_x_dy	0.08	0.06	0.09	0.06	0.05				0.13	0.46	1.49	1.55	2.06	2.00	1.55	1.44	1.36	0.50	0.05	0.09
betasq_x_mom36m	−0.02	−0.30	−0.09	−0.12	−0.22	−0.35	−0.35	−0.04	−0.05	−0.25	−0.12	−0.04	−0.05							
indmom_x_mom12m	−0.13	−0.22	−0.14	−0.03						−0.04	−0.02	−0.02	−0.01	−0.24	−0.37	−0.42	−0.27	−0.15	−0.06	−0.10

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
dolvol_x_idiovol	−0.52	−0.41	−0.53	−0.88	−1.07	−1.09	−1.16	−1.16	−1.04	−0.93	−0.82	−0.66	−0.64	−0.67	−0.52	−0.60	−0.65	−0.52	−0.38	−0.06	
ill_x_mve0	−0.47	−0.65	−0.71	−0.58	−0.58	−0.58	−0.47	−0.39	−0.27	−0.15	−0.18	−0.07								−0.09	
beta_x_lev	−0.79	−0.71	−1.70	−1.71	−1.61	−1.42	−1.30	−0.87	−0.44						−0.03	−0.02	−0.01				
dolvol_x_zerotrade	0.07	0.59	1.65	1.65	1.78	1.83	1.76	1.67	1.20						0.03	0.01	0.01	0.04	0.02	0.02	
mom1m_x_turn	0.14	0.09	0.11	0.13	0.14	0.13	0.10	0.08	0.05	0.08	0.13	0.04	0.14			0.02					
beta_x_std_dolvol	−0.18	−0.07	−0.07	−0.08	−0.06	−0.11	−0.14	−0.09							−0.02	−0.01	−0.02	−0.05	−0.03		
mom12m_x_mom1m															−0.03	−0.08	−0.08		−0.02	−0.05	
betasq_x_mom1m												0.12	0.09	0.17	0.10	0.04	0.05	0.02	0.05		
mvel1_x_zerotrade	0.38	0.16							0.60	1.88	2.58	2.71	2.28	2.61	2.14	1.91	2.16	1.92	1.99	0.75	
beta_x_mom1m	0.03											0.02									
mom1m_x_mom6m	−0.02			−0.02	−0.02	−0.01	−0.03	−0.02	−0.01			−0.02	−0.09	−0.06	−0.05					−0.05	
dy_x_ps	0.23	0.05	0.69	0.35	0.20	0.07															
indmom_x_mom1m										−0.08	−0.01					−0.01	−0.03	−0.11	−0.15	−0.18	
age_x_dy	0.45	0.40	0.10	0.53	0.33	0.24	0.19	0.01													
indmom_x_mom36m											0.10	0.17	0.20	0.20	0.34	0.36	0.34	0.15	0.02		
gma_x_lev					−0.02			−0.05				−0.17	−0.21	−0.26	−0.31	−0.16	−0.03	−0.20	−0.21	−0.06	
dy_x_ms	0.04	0.02																			
indmom_x_maxret	0.07																				
chmom_x_mom1m																					
cashdebt_x_dy	0.06																				
betasq_x_mom36m				−0.07															−0.02		
indmom_x_mom12m	−0.03																				

Figure 15: $L2$-**boosting coefficients: first-order effects**

The figure plots the estimated coefficients of the feature's first-order effects when using the $L2$-boosting method across all estimation rounds.

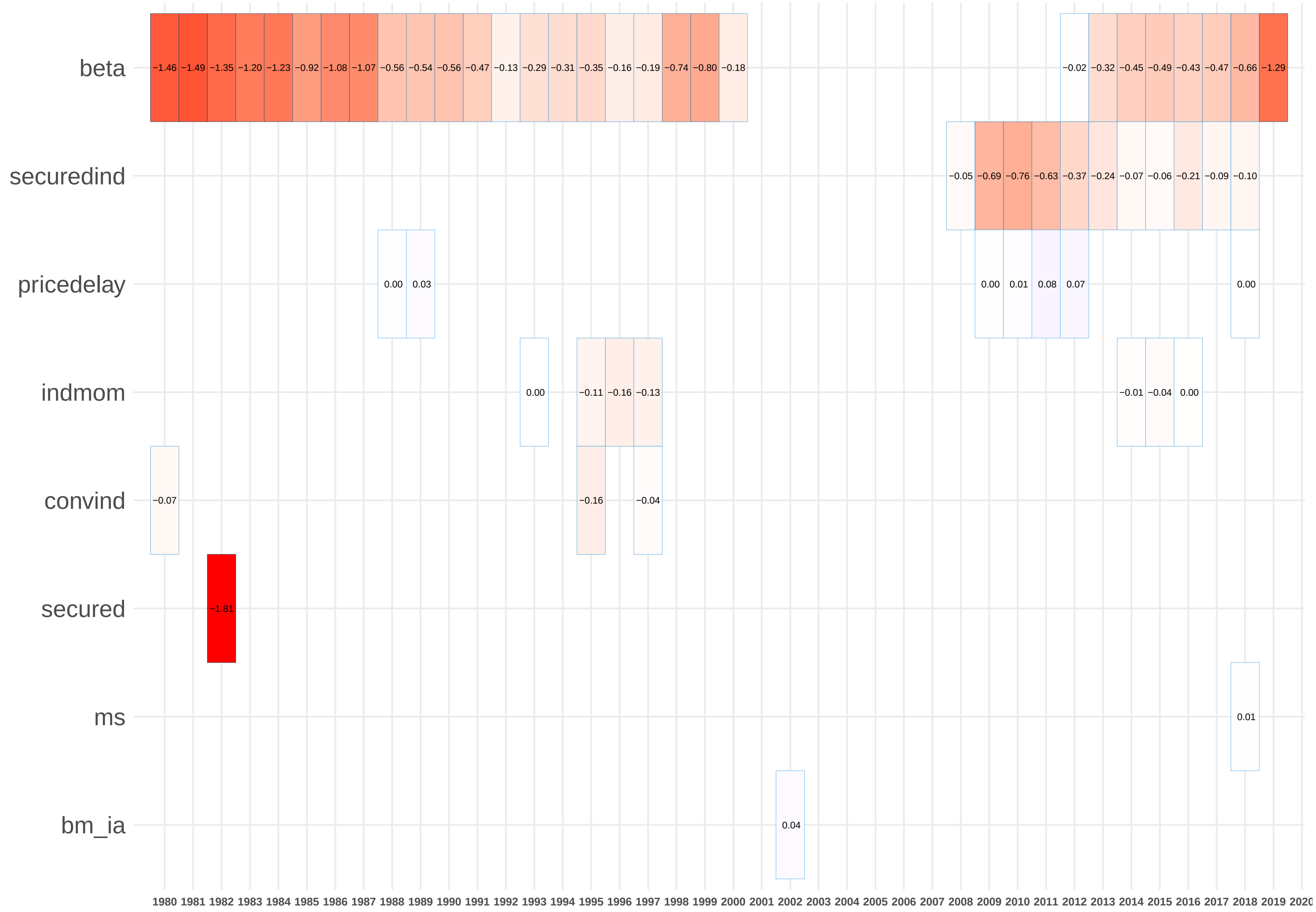


Figure 16: $L2$-boosting coefficients: second-order effects

The figure plots the estimated coefficients of the feature's second-order effects when using the $L2$-boosting method across all estimation rounds.

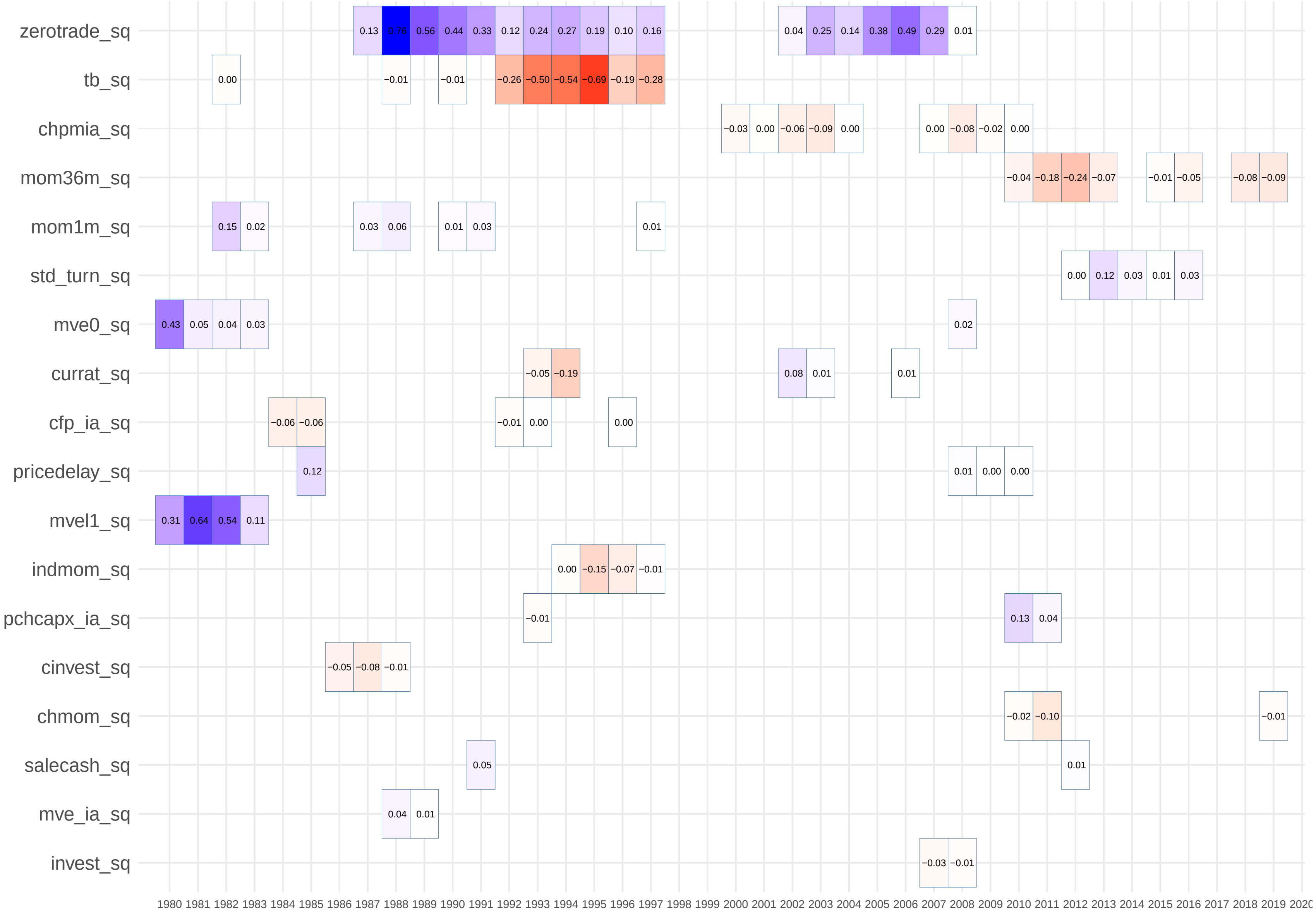
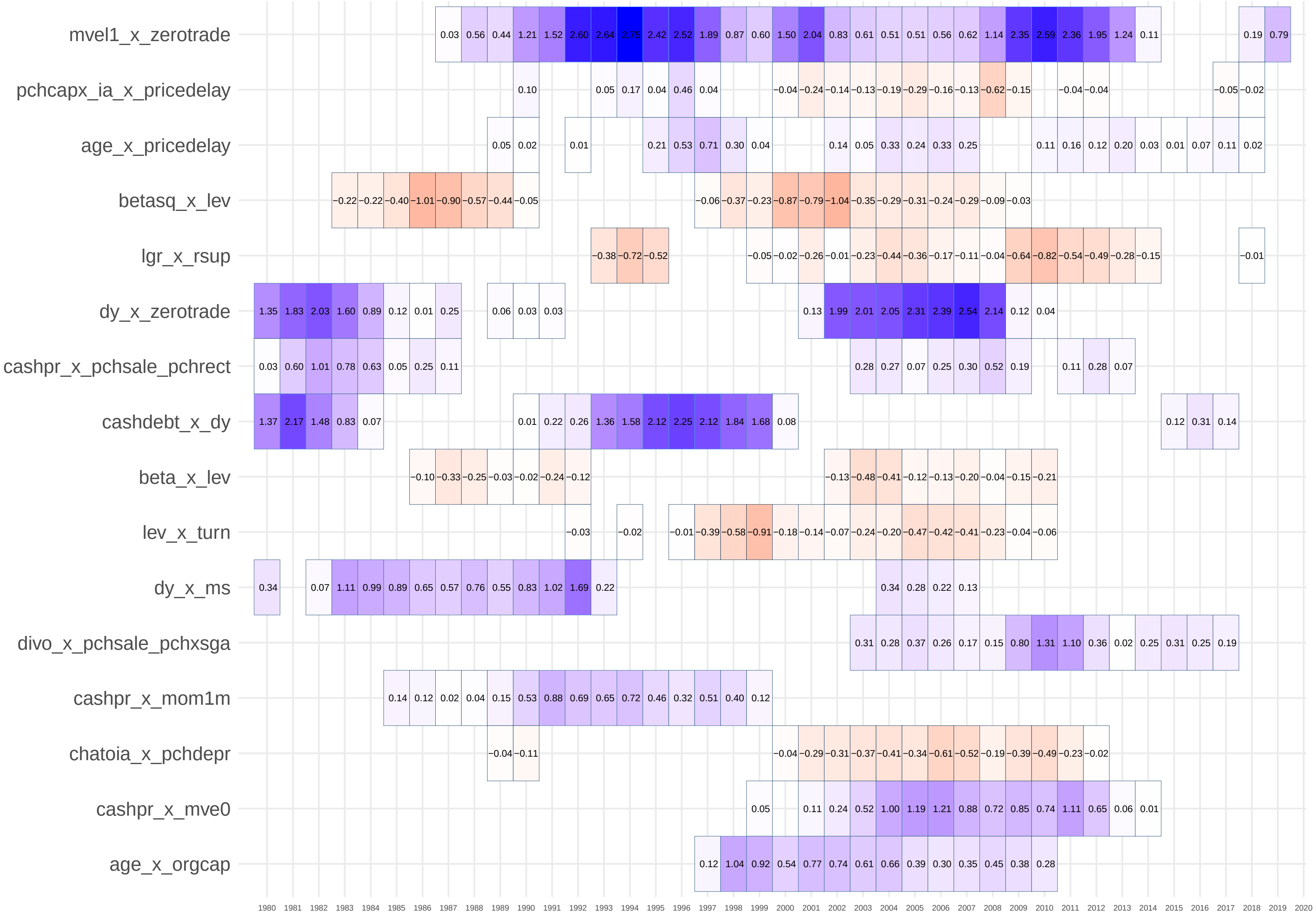


Figure 17: $L2$-boosting coefficients: interactions

The figure plots the estimated coefficients of the feature's interactions when using the $L2$-boosting method across all estimation rounds.



Algorithms

The $L2$-boosting algorithm commences by initializing the parameter vector, $\hat{\theta}^{[0]}$, to zero and setting the initial portfolio weights, $w_t(\hat{\theta}_0)$, equal to the benchmark weights, w_{b_t}. The algorithm then enters a loop that continues until a predefined stopping iteration m^* is reached. In each iteration m, the algorithm systematically goes through each of the K characteristics. After evaluating all K characteristics, the algorithm selects the characteristic, k^*, that minimizes the empirical objective function in eq. (4). It then updates the parameter vector $\hat{\theta}^{[m]}$ to be zero for all elements except for the k^*-th element, which is set to $\hat{\theta}_{k^*}^{[m]}$. The algorithm also updates the parameter vector $\hat{\theta}_m$ using a step size $\nu \in (0, 1)$, and accordingly adjusts the portfolio weights, $w_t(\hat{\theta}_m)$, based on these updated parameters. The iteration counter m is incremented, and the process repeats until the termination criterion is met.

Algorithm 1 $L2$-boosting for minimum-variance parametric portfolios

```
Initialize $\hat{\theta}^{[0]} = 0$, $w_t(\hat{\theta}_0) = w_{b_t}$ and $r_{b_0} = w_t(\hat{\theta}_0)^\intercal r_{t+1}$
Set $m = 1$
while $m \leq m^*$ do
    for $k = 1$ to $K$ do
        Solve the empirical investor problem for the $k$-th characteristic:
        Minimize the objective function
            $\frac{1}{2}\frac{1}{(T-1)}\sum_{t=1}^{T-1}\left(\dot{r}_{b_m} + \theta_k^{[m]}\dot{r}_{c_k,t+1}\right)^2$
        using the estimator:
            $\hat{\theta}_k^{[m]} = -\frac{\hat{\sigma}_{b_{m-1}c_k}}{\hat{\sigma}_{c_k}^2}$
    end for
    Choose $k^*$ that minimizes the objective function for all $k$
    Set $\hat{\theta}^{[m]}$ to zero except for the $k^*$-th element, which is $\hat{\theta}_{k^*}^{[m]}$
    Update $\hat{\theta}_m$ using a step size $\nu$:
        $\hat{\theta}_m = \hat{\theta}_{m-1} + \nu\hat{\theta}^{[m]}, m \geq 1$
    Update $w_t(\hat{\theta}_m)$:
        $w_t(\hat{\theta}_m) = w_{b_t} + \hat{\theta}_m^\intercal X_t / N_t,\ m \geq 1$
    Update $r_{b_m}$ :
        $r_{b_m} = w_t(\hat{\theta}_m)^\intercal r_{t+1}$
    $m = m + 1$
end while
```

Algorithm 2 Marginal Screening

```
Require: Design matrix $X \in \mathbb{R}^{n \times p}$, response vector $y \in \mathbb{R}^n$, model size $k$
Ensure: Estimated coefficients $\hat{\beta}_{\hat{S}}$ for the selected variables
Compute the product $X^T y$ to obtain a vector in $\mathbb{R}^p$
Compute the absolute values $\left|X^T y\right|$ to get the marginal correlations
Identify $\hat{S}$, the index set of the $k$ largest entries in $\left|X^T y\right|$
Extract the submatrix $X_{\hat{S}}$ from $X$ using the indices in $\hat{S}$
Compute $\hat{\beta}_{\hat{S}} = \left(X_{\hat{S}}^T X_{\hat{S}}\right)^{-1} X_{\hat{S}}^T y$ to estimate the coefficients for the selected variables.
return $\hat{\beta}_{\hat{S}}$
```