Variable selection for minimum-variance portfolios*

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

July 20, 2024

Abstract

Machine learning (ML) methods have been successfully employed in identifying variables that can predict the *equity premium* of individual stocks. In this paper, we investigate whether ML can also be helpful in selecting variables relevant for *optimal portfolio choice*. To address this question, we parameterize minimum-variance portfolio weights as a function of a large pool of firm-level characteristics, including their second-order and cross-product transformations, yielding a total of 4,610 predictors. Our findings indicate that employing ML to select relevant predictors significantly enhances portfolio performance: minimum-variance portfolios achieve lower risk compared to sparse specifications commonly considered in the literature, especially when non-linear terms are included. Moreover, some of the selected predictors not only reduce portfolio risk but also enhance returns, resulting in minimum-variance portfolios with strong performance in terms of Sharpe ratios in certain situations. Specifically, the L2-boosting method emerged as the most effective approach, identifying predictors that minimize risk, reduce covariance with the benchmark, and increase risk-adjusted returns. Our evidence suggests that ad-hoc sparsity can be detrimental to the performance of characteristics-based portfolios.

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

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, and the Economics of Financial Technology Conference 2024, 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 to select them? 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, **????**). 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 **?**. In this paper, we assemble a data set with 4,610 firm-level predictors and demonstrate that machine learning (ML) variable selection methods are able to select 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 few predetermined state variables in order to explain the cross-section patterns of stock returns (e.g. **???**) and to determine how they affect optimal portfolio allocations (**?**). **?** refer to this approach as "characteristics-sparse models". For instance, **?** and **?** propose porfolio allocation methods based on low-dimension 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 **?** and **?** 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 few ad-hoc characteristics. **?** 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 put forward a methodological framework for high-dimensional predictor selection and provide empirical evidence that allows us to establish four main conclusions that are useful for academics and market practitioners. First and foremost, the choice of the variable selection method matters: there are substantial differences in terms of the *preference* of each method to specific features and to the *number* of features each method selects. Second,

predictors based on transformations of the original characteristics (especially *interactions* between characteristics) are selected more often than the predictors' first-order (main) effects, indicating that accounting for non-linear effects is important for portfolio construction.[1] Third, the number of selected features that matter for the portfolio construction problem exceeds the number of variables used in popular low-dimension factor models. Fourth, working with a subset of important predictor variables translates into minimum-variance portfolios that outperform their benchmarks in terms of lower standard deviation of out-of-sample portfolio returns.

We add to the existing literature by extending the works of **?** and **?** along two dimensions. First, we raise the statistical challenge by assembling a set of predictor variables containing the 95 firm characteristics considered in **?** 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 matter for portfolio construction, but also *what type* of non-linearity (i.e. non-linearities of individual predictors or interaction between covariates) is most important in order to help the investor improve portfolio performance. Existing empirical evidence (**??????**) 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 and what type of non-linearities are relevant *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 alternative approaches to variable selection: *regularization*, *ensembles*, and *Bayesian* methods. While regularization methods perform variable selection by shrinking the coefficients using the $L1$ regularization (or a combination of $L1$ and $L2$ regularizations), the ensemble method considered in the paper operates via a stagewise additive modeling procedure which 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

[1]Non-linear transformations and interactions between firm characteristics have been extensively studied in the asset pricing literature. For instance, **?** assume that the stochastic discount factor is quadratic in the market return. **?** 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. **?** 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.

remain unshrunk. Finally, as a robustness check, we implement the marginal screening method of **?**, which is computationally efficient and suitable for ultra-high cross-sections of predictors.

It is also worth highlighting three important methodological aspects of our work. First, consistent with our interest on 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 **?** 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 **?** to parameterize portfolio weights as a linear function of non-linear transformations of lagged firm characteristics. We borrow analytical results obtained in **?** 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 the most popular smart-beta strategies in the U.S. (**?**).[3] Second, minimum-variance portfolios are less sensitive to estimation errors in the expected returns, and often leads to robust performance in practice (**?**). Third, the minimum-variance parametric portfolios can be cast as an unconstrained regression problem, which allows deploying 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

[2] **?** 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. **?** builds on the approach developed by **?** and find that, although the 4-factor model of **?** is better than the 5-factor model of **?** at explaining anomalies, the two models have similar performance in terms of investing.

[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`).

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 matter 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 specification such the 3- and 5-factor models of **??** as well as classic benchmark strategies such as the value-weighted and equally-weighted portfolios. For instance, the minimum-variance portfolios with predictors selected with the best performing method delivered out-of-sample returns with 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-factors (14.6%) models.

Our empirical application reveals that the alternative feature selection methods converge in highlighting the dominance of interaction terms within the subset of selected features. On average, interactions correspond to more than 85% of the selected features. These results not only highlights the importance of explicitly accounting for non-linearities in the portfolio construction problem, but also confirms that interactions should be a more plausible source of non-linearity in comparison to high-order moments, as highlighted in **?**. The alternative methods, however, disagree on the desired level of sparsity: the average number of selected predictors varies between 33 and 75 features across methods.

We also investigate the importance of individual predictors. We find that 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 interpretation of the predictor importance by i) identifying which aspects of the investor's utility the predictors contribute the most, and ii) showing how predictor importance translates to

portfolio allocations. For instance, stocks with lower liquidity get higher weights, and this result is stronger for stocks with lower idiosyncratic volatility. Stocks with lower market betas get higher weights, and this result is stronger for stocks with higher standard deviation of liquidity. Our results are in line with those reported in **?**, who 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.

Our paper is connected to a growing literature that employs machine-learning methods to support portfolio choice decisions; see **?** for a recent review of the literature. **?**, **?**, and **?** employ ML methods to construct prediction-based portfolios of mutual funds, whereas **?** focus on prediction-based stock portfolios. ML methods have also been employed in economic and financial forecasting. For instance, **?**, **?** and **?** employ the lasso and the elastic-net to select regressors for stock return prediction, whereas **?** employ ML to predict the aggregate market excess return. **?**, **?**, **?**, and **?** employ the boosting method for macroeconomic forecasts. A common denominator to these studies is that employing variable selection methods lead to improved results in comparison to those obtained with standard statistical methods.

Our paper is mostly connected with the works of **?** and **?**. **?** formulate 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 **?** can capture non-linearities in a flexible way, our method is highly interpretable and allows us to understand not only *how many* predictors survive the different variable selection processes but also *why* a selected predictor is important according to a given method. **?** use 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, focus 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 for portfolio construction.

The rest of the paper is organized as follows. Section 2 describes the data set used in the paper. Section 3 detailed 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 data set of 95 monthly firm characteristics used in **?**. The data base was downloaded from Dacheng Xiu's homepage (`https://dachxiu.chicagobooth.edu/download/datashare.zip`).[4] Stock returns come from CRSP. Our sample spans from January 1973 until June 2019 (574 months).

We perform feature engineering to augment the original data set containing the 95 predictors' main effects. Specifically, for each firm characteristic, we obtain the square values and cross-products among all predictors, that is, we define $pred_{i,t}^2$ and $pred_{i,t} \times pred_{j,t}$ for $i \neq j$ to denote the second-order and interaction effects, respectively.[5] The augmented data set 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 **?**. We also follow **?** and standardize each predictor so that its cross-sectional mean is zero and standard deviation is one. Missing characteristic values are set to zero. The resultant standardized predictor is a long-short portfolio that goes long stocks whose characteristic is above the cross-sectional average, and short stocks whose characteristic is below the cross-sectional average.

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

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

[4]Table A.6 of the supplementary material of **?** 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 **?** 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.

[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*.

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 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) 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 a 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 **?** and its minimum-variance reformulation as an unconstrained regression problem. Second, we discuss how the alternative variable selection methods can be introduced into that 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 (like the 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 asset.

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 up 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 in portfolio optimization considers the minimum-variance policy; see, for instance, **?**, **?**, **?**, **?**, and **?** 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 that solves the following problem:

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

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 regression problem. Specifically, the optimal parameter θ^* in eq. (4) can be estimated with a time-series regression without 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.

3.2 Variable selection methods

We now discuss the three classes of variable selection methods that helps solving the empirical minimum-variance parametric portfolio problem in eq. (4) in situations in which the number of predictors can exceed the number of data points. Three major classes of variable selection procedures are considered: 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 **?** 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

[6]The regression formulation of the parametric minimum-variance portfolio problem in (5) differs from existing portfolio-regression formulations. For instance, **?** shows that the tangency portfolio can be obtained by a regression of a constant **l** onto a set of asset's excess returns, without an intercept term. **?** 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 **?** as it is based on the parametric-portfolio formulation.

coefficients. The ridge penalization is defined as

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

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, **?**).

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

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

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, **?**).

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. **?** proposed the adalasso regression, which is a modified version of the lasso regression based on adaptive weights used to penalize coefficients differently,

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

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 (**?**), 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 **?**, the elastic net strategically combines the merits of both lasso and ridge regression, and the penalization function becomes

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

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, **?**).

Smoothly clipped absolute deviation (scad) The scad penalization introduced by **?** 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 **?** 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 collect 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, **?**). A successful global-local shrinkage prior is the horseshoe (**?**), 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 delivered by the scale mixture of normals defined by (14), which shrink weak signals toward the origin, while still allowing strong signals to remain unshrunk via its 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* (**?**) and *boosting* (**??**) are commonly used, where models are trained on various subsets or reweighted instances of data. In this paper, we consider the $L2$-boosting method of **?** which falls under the

umbrella of boosting algorithms. The $L2$-boosting method based on the principle of improving a model's prediction 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, e.g., cannot be applied if K is larger than the number of observations T). Algorithm 1 provides a pseudo code to implement the $L2$-boosting for minimum-variance parametric portfolios.[7]

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 **?**, 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 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 **?** and **?** and employ the corrected AIC criterion to select the optimal number of iterations, m^*,

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

[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.

where df is the effective number of degrees of freedom after m iterations of the boosting algorithm.[8] As m increases, $\hat{\text{var}}\left[r^m_{p,t+1}\right]$ 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 via the efficient algorithm proposed by **?**, and the global-local parameters τ and λ_i are updated using the slice sampler of **?**.

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 **?**, 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 an *individual* predictor. For that purpose,

[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.

we calculate their *marginal contributions* using the first order derivative of eq. (16),

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

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 to 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 own-variance, $\text{diag}(\hat{\Sigma}_c)\theta$, the predictor covariance with other predictors, $(\hat{\Sigma}_c - \text{diag}(\hat{\Sigma}_c))\theta$, 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 assuming a equally-weighted (EW) benchmark portfolio. Two reasons motivate our choice for this benchmark portfolio. First, the well-documented good performance of this strategy relative to more sophisticated ones, as extensively discussed in **?**. Second, **?** show that the EW strategy emerges naturally in a mean-variance context since an optimal mean-variance portfolio can be 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 when using only the original 95 predictors used in **??**. Second, we use the augmented set of 4,610 predictors which includes the second-order and cross-products transformations. This allows us to understand the extent to which non-linear transformations of firm characteristics help improving portfolio performance.

Given our interest in the minimum-variance strategy, the most relevant performance metric

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

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 the 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 $t = T, \ldots, T_{tot} - 1$ and for each strategy j. Holding the portfolio w_t^j for one month gives the out-of-sample return net of transaction costs 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

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

where e_{t-1} is the N_{t-1} dimensional vector of ones and $x \circ y$ is the elementwise 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:

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

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 denotes 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 SR of two

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

portfolios by using the two-sided p-value of the prewhitened HAC_{PW} test described by **?** and **?** for the portfolio variance and the SR, respectively. Specifically, we test for differences in portfolio variances and SR of the various strategies with respect to those of the minimum-variance parametric portfolio obtained when all 4,610 predictors are used along 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 on the parameters in which all coefficients are set to zero. Second, we implement the minimum-variance parametric portfolios using the characteristics from the 3-factor (market, size, and value) and 5-factor (market, size, value, profitability, and investment) models of **?** and **?**, 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 results. First and foremost, the choice of the variable selection approach matters: the different methods lead to minimum-variance parametric portfolio with different performance. We observe that the lasso method leads minimum-variance portfolios with lower portfolio risk in comparison to all other methods for both configurations of the data set. Three methods emerge as the best performers in terms of 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 the 4,610 predictors and the lasso method are used (8.3%). This figure, however, is statistically indistinguishable relative to 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 with the OLS method. Specifically, using the characteristics from the popular 3-factor and 5-factor models lead to a portfolio standard deviation of 15% whereas using the OLS method leads a portolio 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 compared to 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 being 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 yield 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 a 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 per Table 2. This suggests that differences in the Sharpe ratio are 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 plots 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 the 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

[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.

($mve0$) 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 influenced significantly 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 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 re-balancing.

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 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 selected 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 obtained with variable selection methods, 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. First, 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. Second, interactions emerged as the most significant source of non-linearity, consistently dominating the selected predictor sets: interaction terms corresponds to 80% to 97% of the selected predictors. This finding aligns with the suggestion of **?** that interactions are likely to provide a more plausible explanation for non-linearity compared to high-order moments. Moreover, 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.

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 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*),

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

and maximum daily return (*maxret*) are selected by lasso and elastic net mainly for their ability to improve investor's utility because of their covariance with the benchmark portfolio. These results align with those reported in **?**, who argues that risk-based characteristics are important drivers of the minimum-variance portfolio allocations. The 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*) is important because it covaries with the benchmark portfolio, whereas square values of short-term reversals (*mom1m_sq*) matters because it has low covariance with the other predictors.

It is also remarkable that interactions between market beta and standard deviation of liquidity ($beta \times std\_dolvol$), liquidity and idiosyncratic volatility ($dolvol \times idiovol$), and bid-ask spread and liquidity ($baspread \times dolvol$) 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.

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 coefficient 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 6 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 **?**, the presence of dominant first principal component (or factor) would result in extreme negative weights in minimum-variance portfolios. Our results seems to contradict this finding as they indicate absence of extreme weighting. One plausible explanation for our results is that our methodology involves constructing minimum-variance portfolio using a *multitude* of characteristics-based factors. In this situation, **?** and **?** show that exploiting multiple factors leads to trading diversification (i.e. netting of trades across factors), which helps reducing extreme positions in individual stocks.

4.4 Screening predictors

The results reported in Table 3 show that selecting few predictors ad-hoc, as in the popular 3-factor and 5-factor models, lead 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 **?** 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 exact recovery of true non-zero coefficients under specific sparsity and signal strength conditions (see **?**). Algorithm 2 describes the procedure to implement 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 7 reveal that selecting the only top-3 predictors using MS lead to minimum-variance portfolio returns with annualized standard deviation of 10.7%, which 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 to 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-dimension factor models in terms of lower risk and higher risk-adjusted returns. Our results also suggest that interactions between characteristics are more important than the individual features' main effects when constructing optimal portfolios. 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, minimize 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 interaction between variables in the portfolio construction problem. Third, they suggest that the number of selected features that matter for the portfolio construction problem is larger 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

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 **??** 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 **??** 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 reports the average total number of selected predictors and their distribution across predictor classes—main effects, second-order effects, and interaction effects—expressed as percentages. All figures represent averages calculated across all estimation rounds.

	Selected predictors	Main effects	Second-order effects	Interaction effects
Lasso	33	6%	7%	87%
Adalasso	62	1%	3%	97%
Elastic net	60	5%	7%	89%
Horseshoe	47	4%	3%	94%
Scad	16	16%	4%	80%
$L2$-boosting	75	1%	2%	97%

Table 6: **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 7: **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 **?** 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.

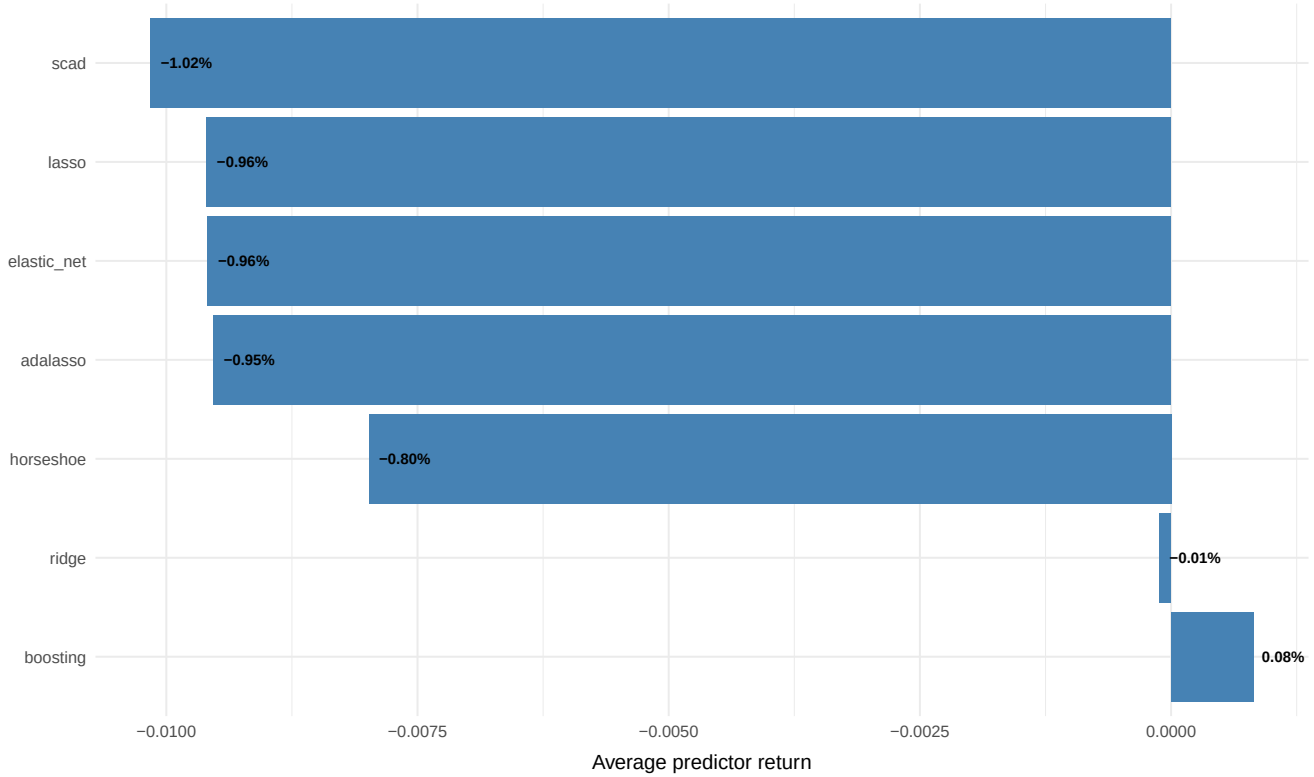


Figure 2: **Individual predictor returns**

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

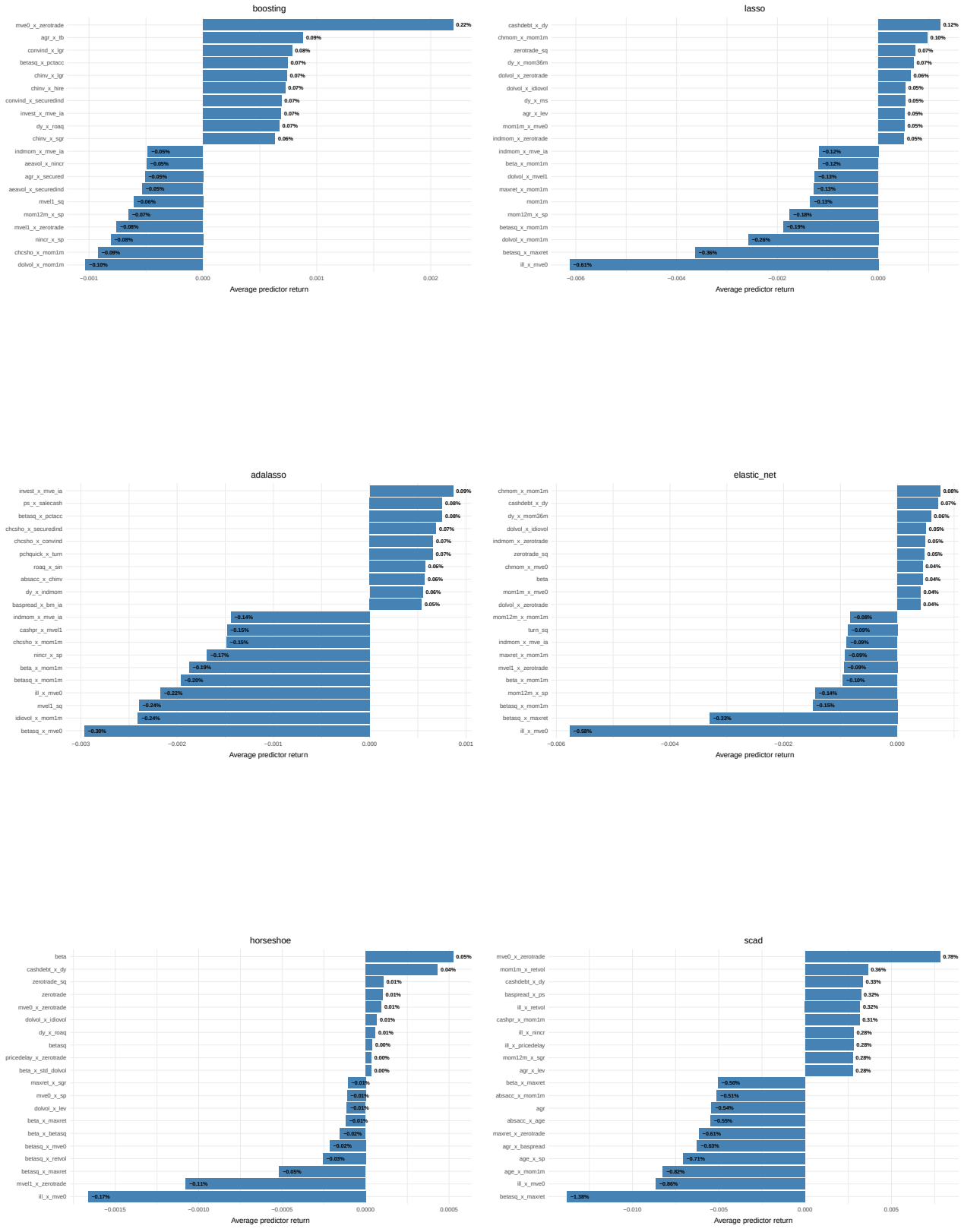


Figure 3: **Lasso predictor importance**

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).

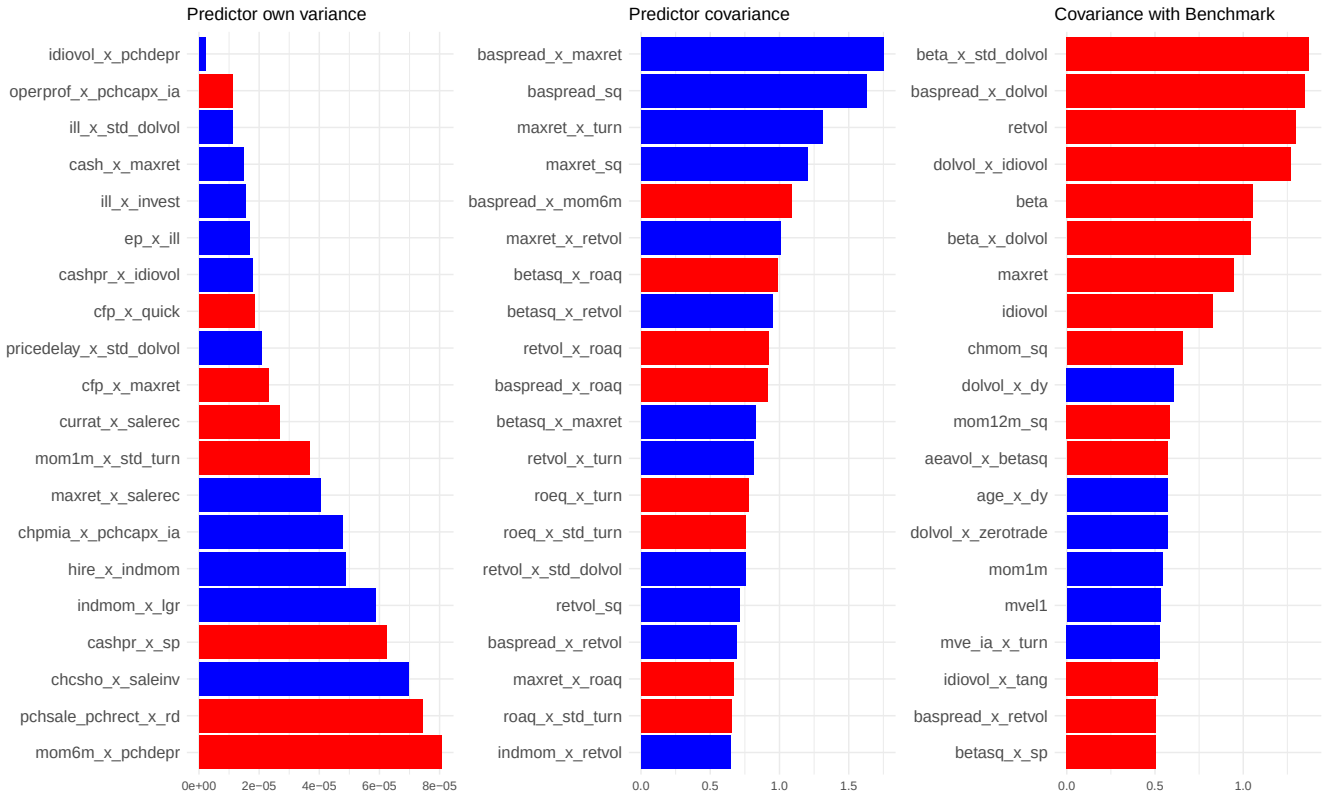


Figure 4: **Elastic net predictor importance**

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).

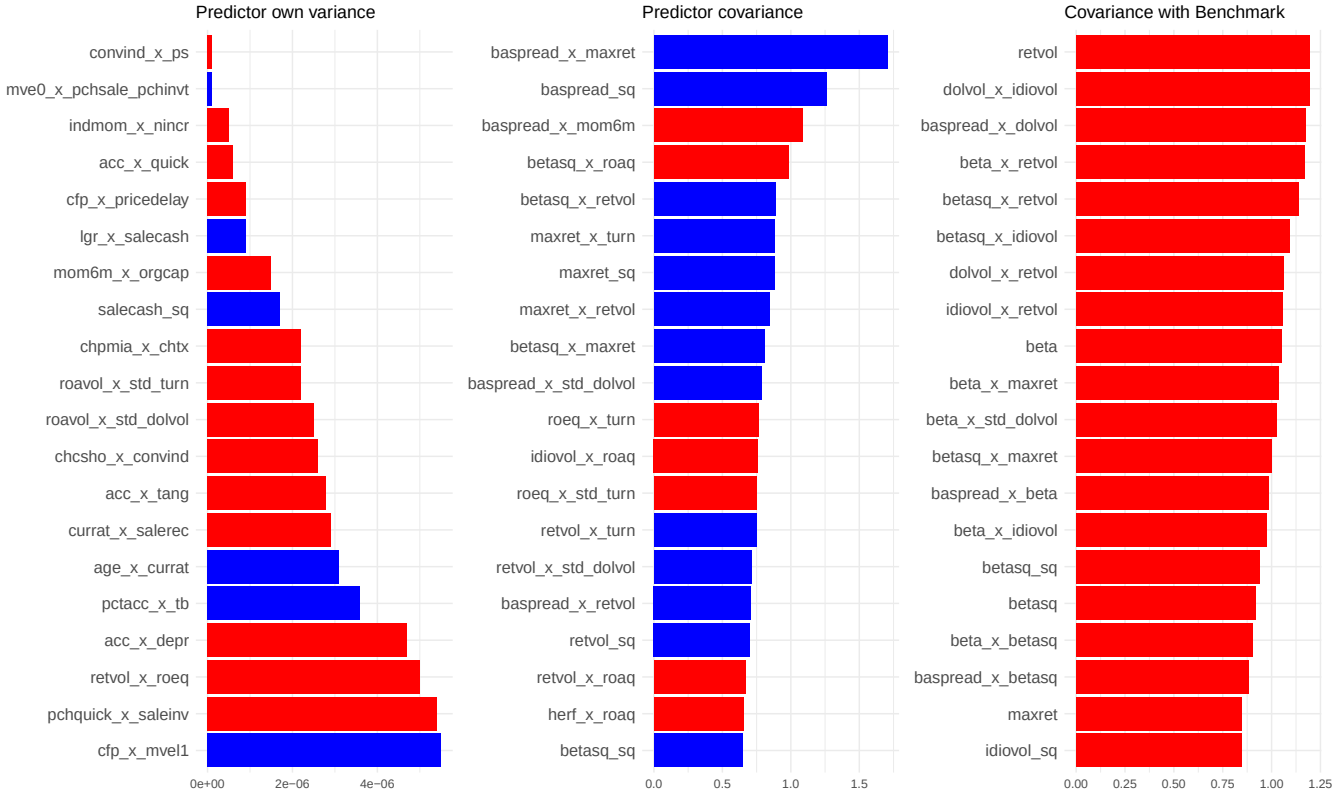


Figure 5: $L2$-boosting predictor importance

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).

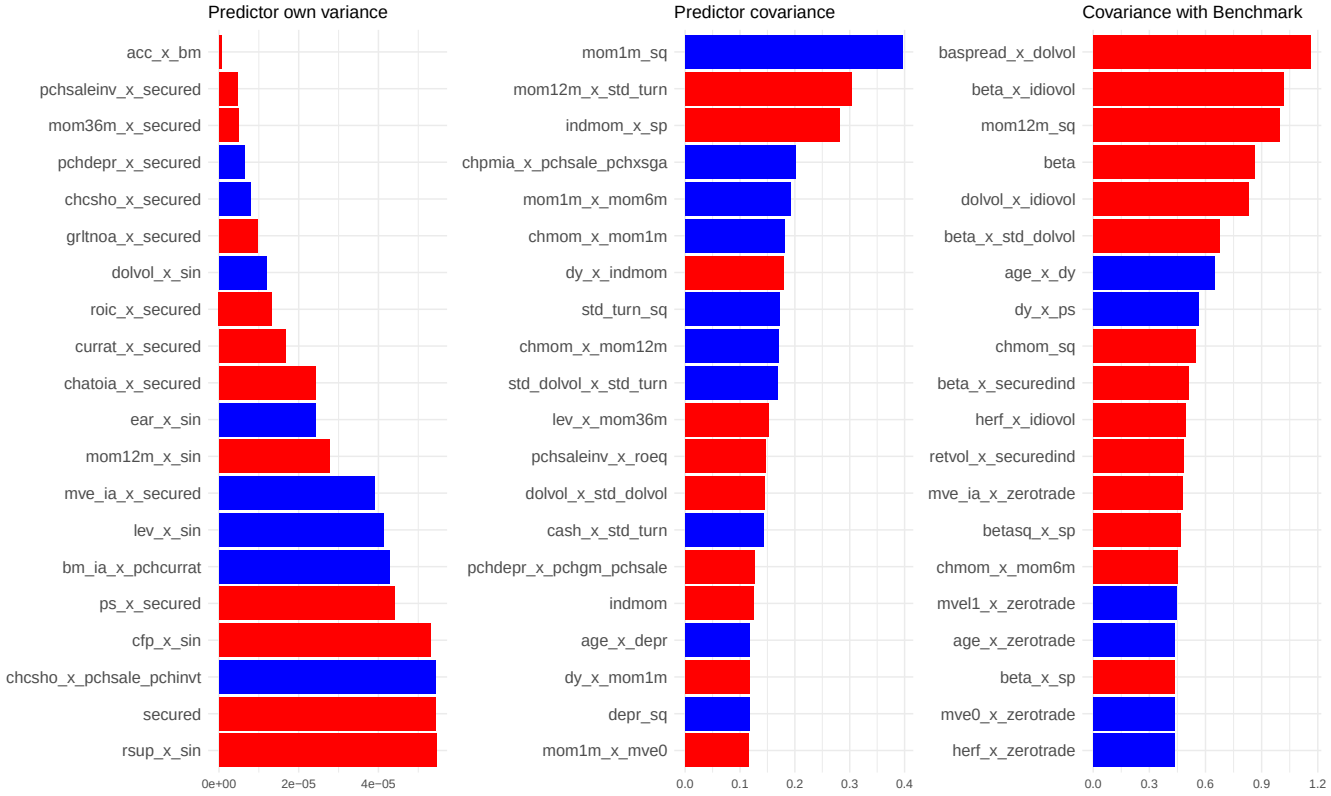
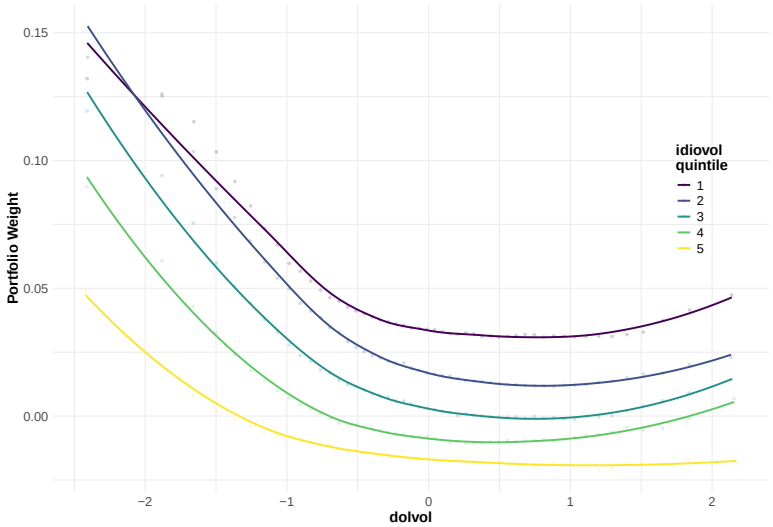
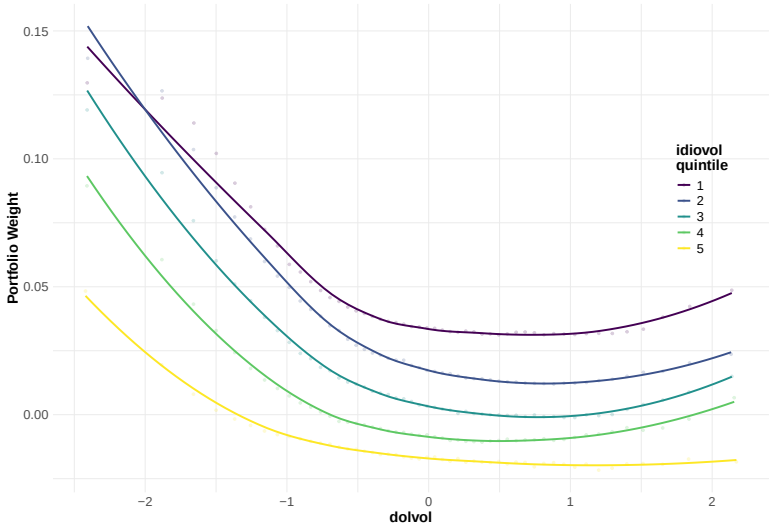


Figure 6: **Portfolio weights, liquidity and idiosyncratic volatility**

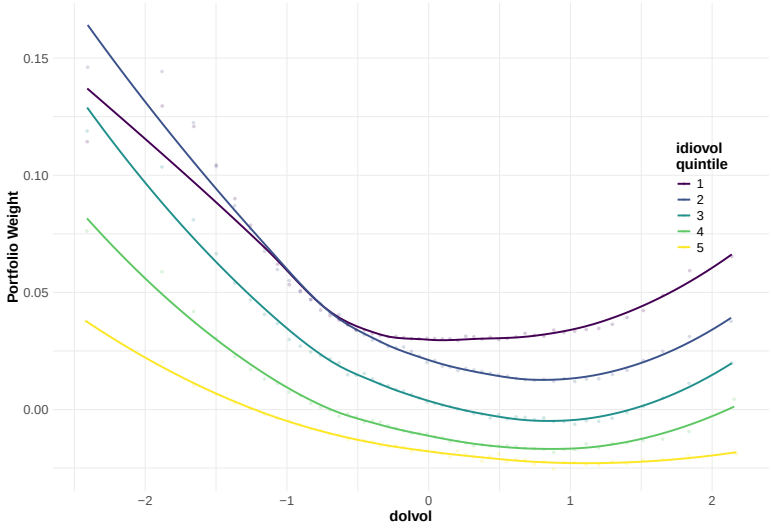
The figure plots the average minimum-variance portfolio weights (vertical axis) across values of the standardized liquidity (*dolvol*) and across quintiles of idiosyncratic volatility (*idiovol*).



(a) Lasso



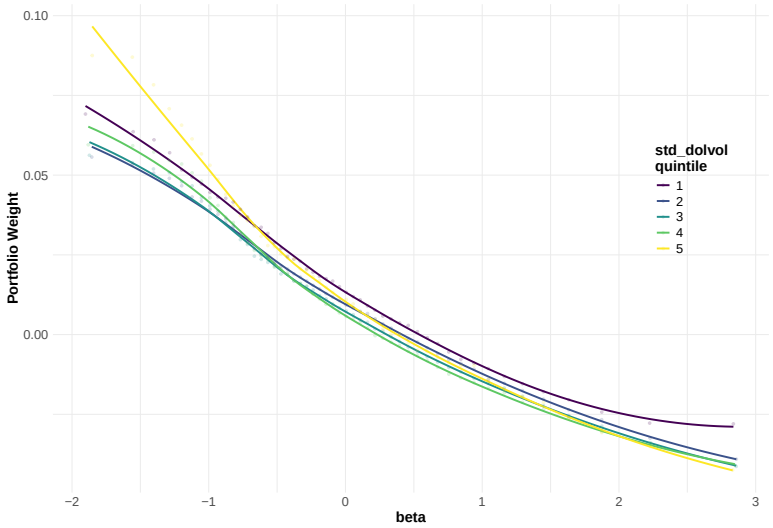
(b) Elastic net



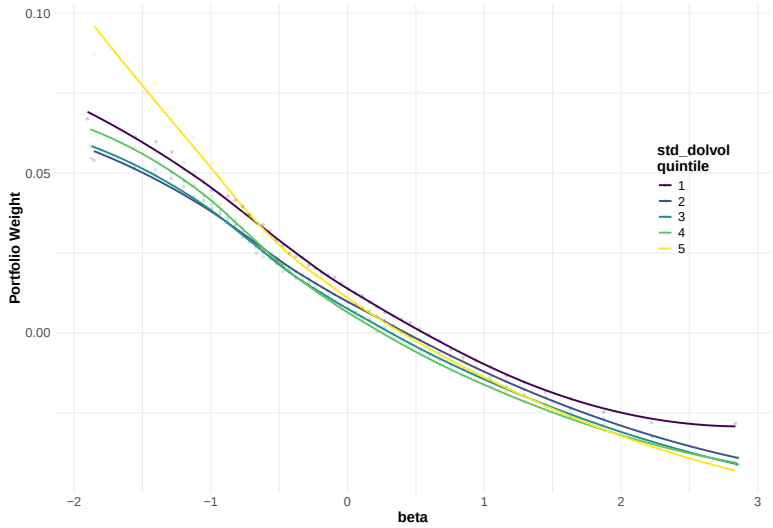
(c) $L2$-boosting

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

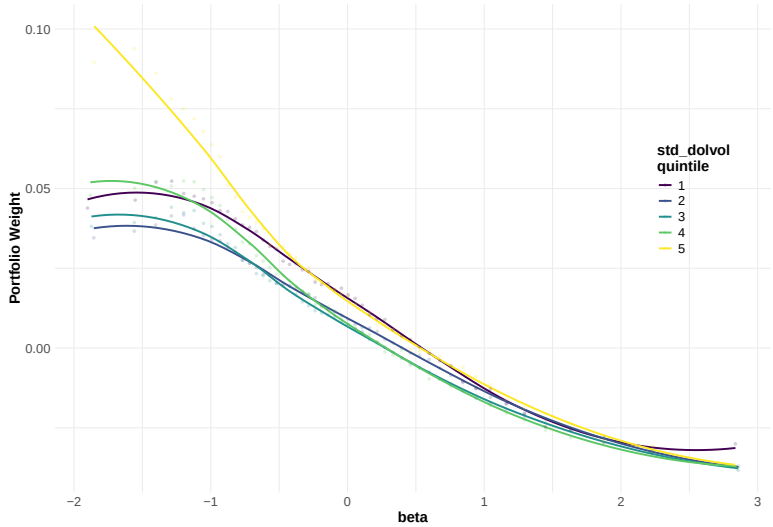
The figure plots the average minimum-variance portfolio weights (vertical axis) across values of the standardized market beta (*beta*) and across quintiles of standard deviation of liquidity (*std_dolvol*).



(a) Lasso



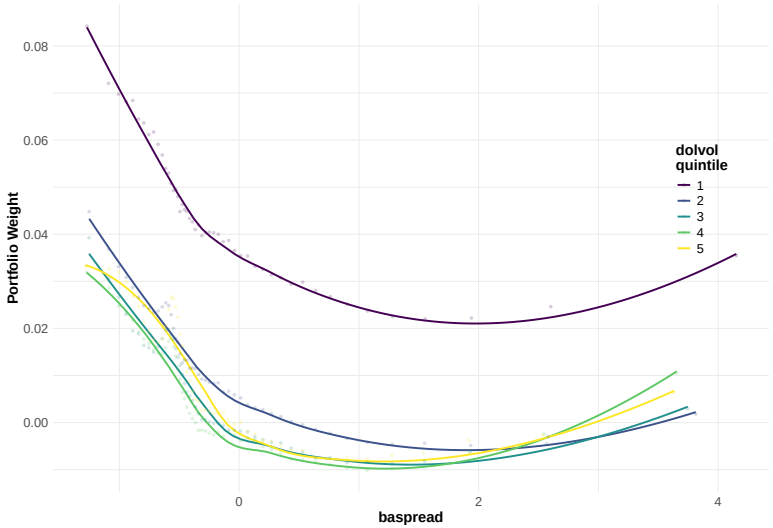
(b) Elastic net



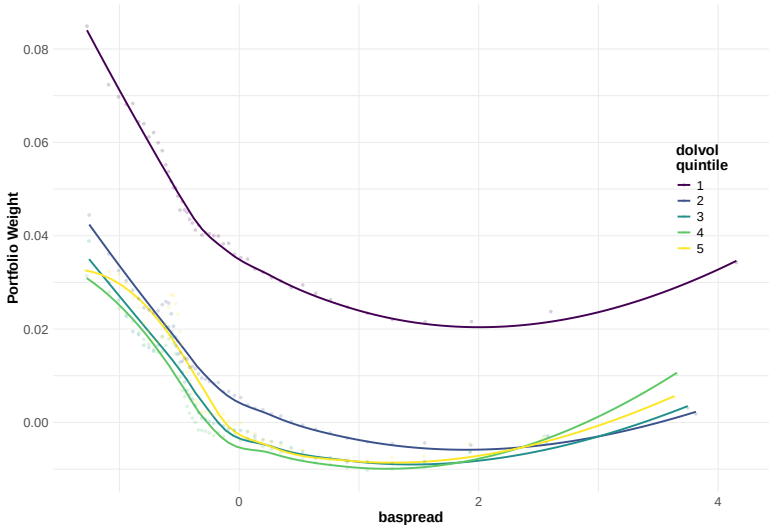
(c) $L2$-boosting

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

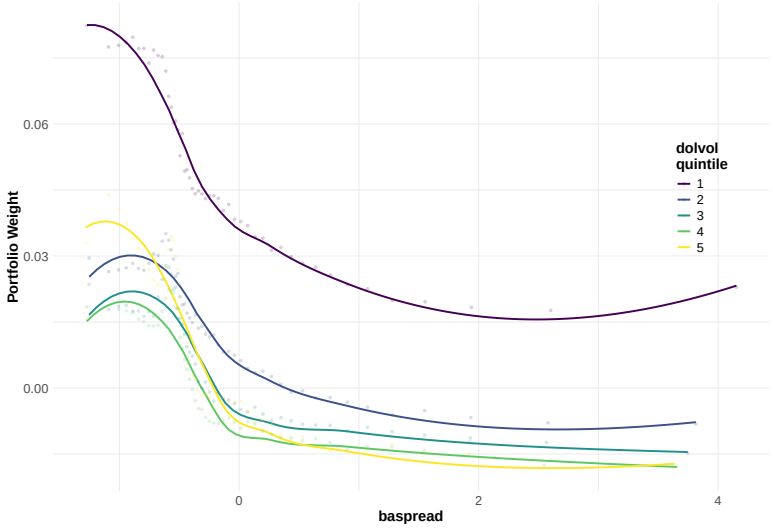
The figure plots the average minimum-variance portfolio weights (vertical axis) across values of the standardized bid-ask spread (*baspread*) and across quintiles of liquidity (*dolvol*).



(a) Lasso



(b) Elastic net



(c) $L2$-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.

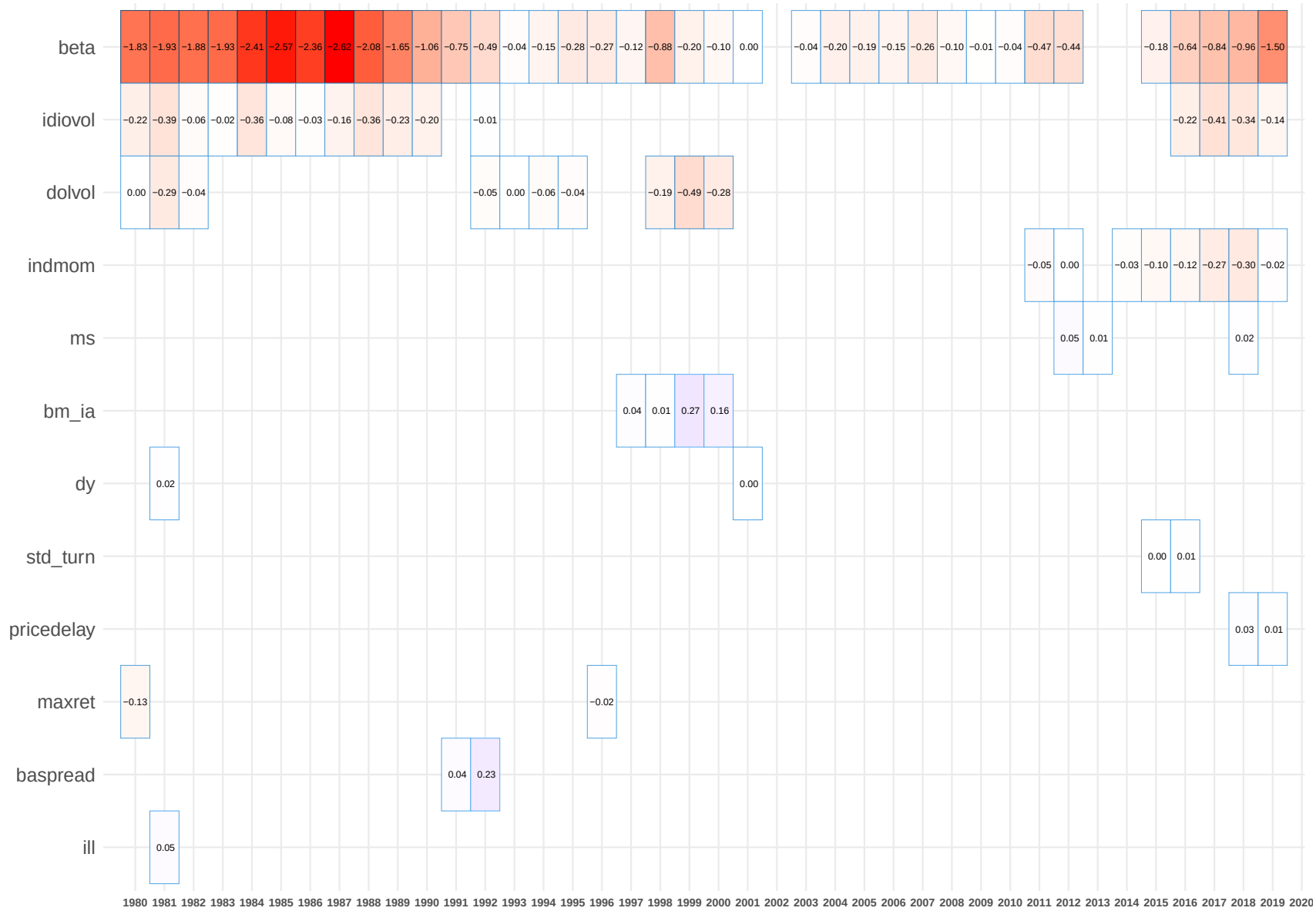


Figure 10: **Lasso coefficients: second-order effects**

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

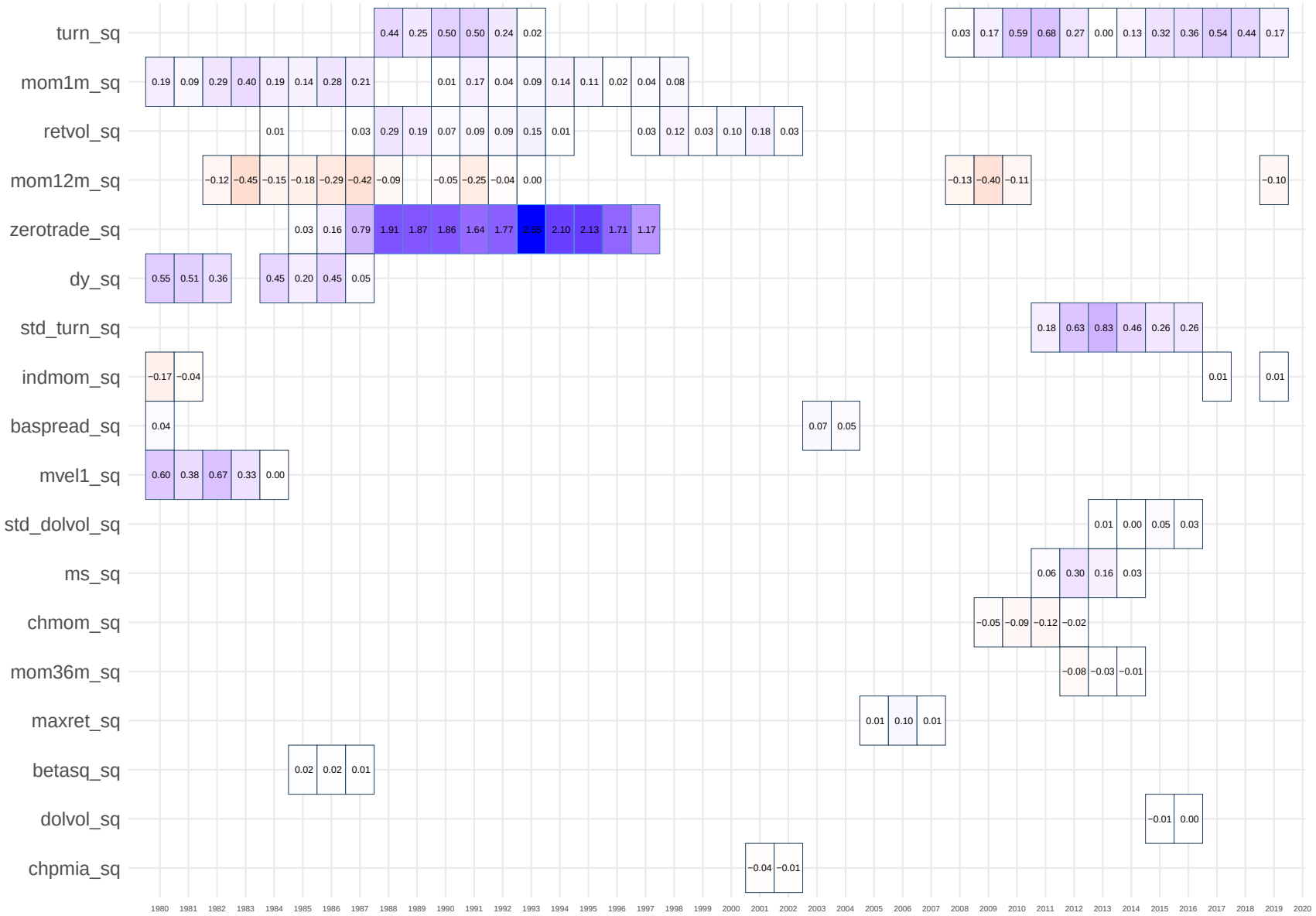


Figure 11: **Lasso coefficients: interactions**

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

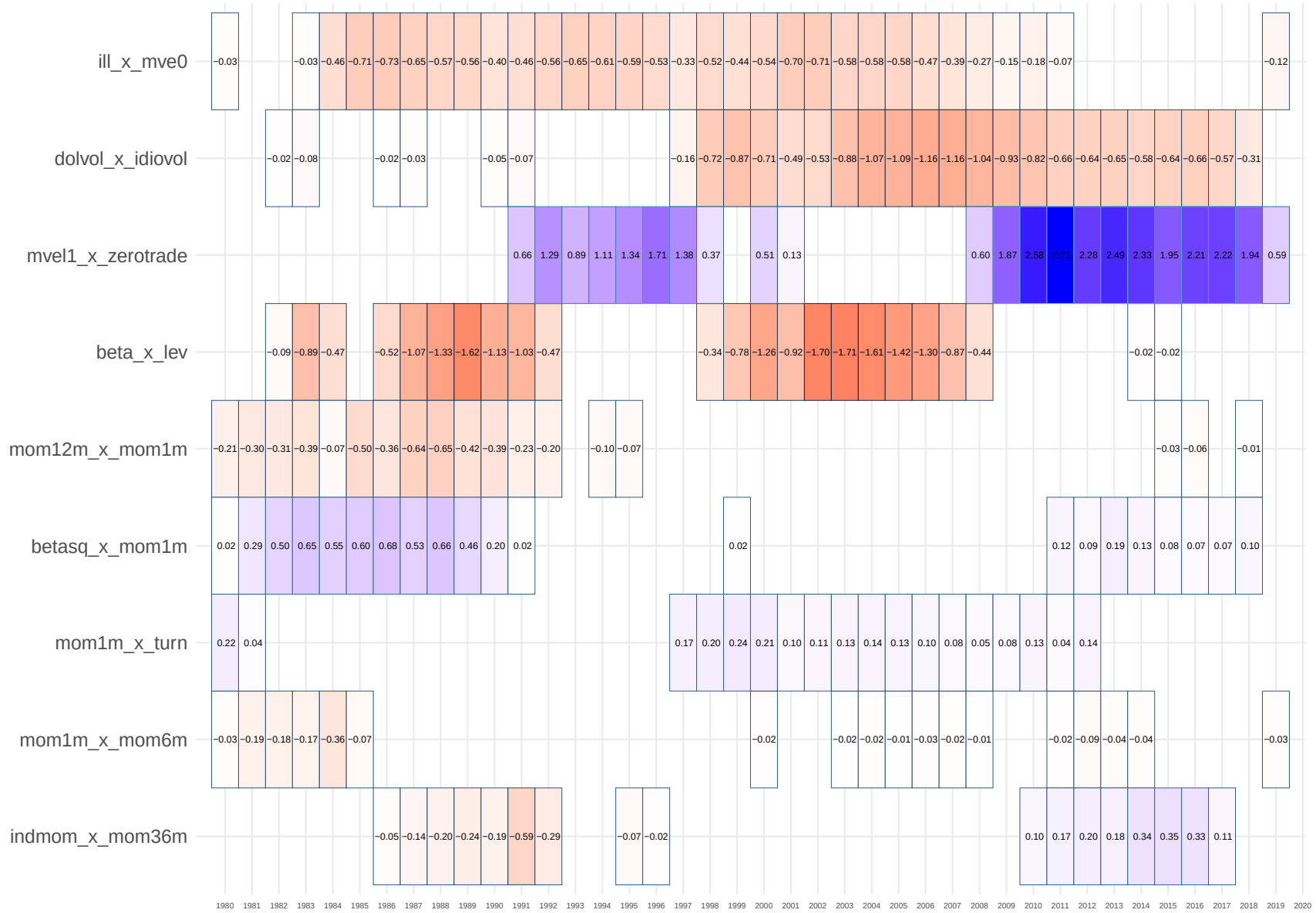


Figure 12: **Elastic net coefficients: first-order effects**

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

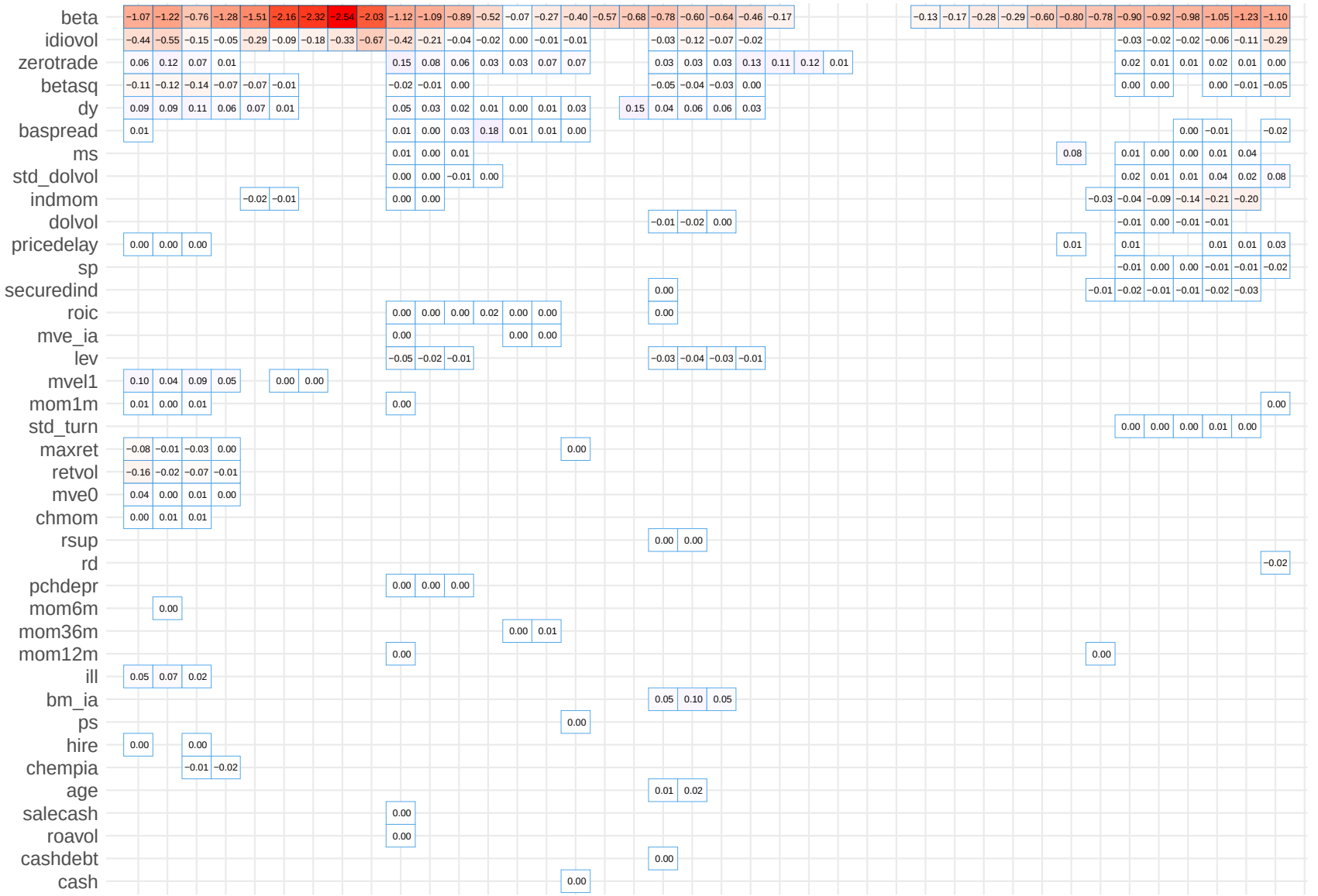


Figure 13: **Elastic net coefficients: second-order effects**

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

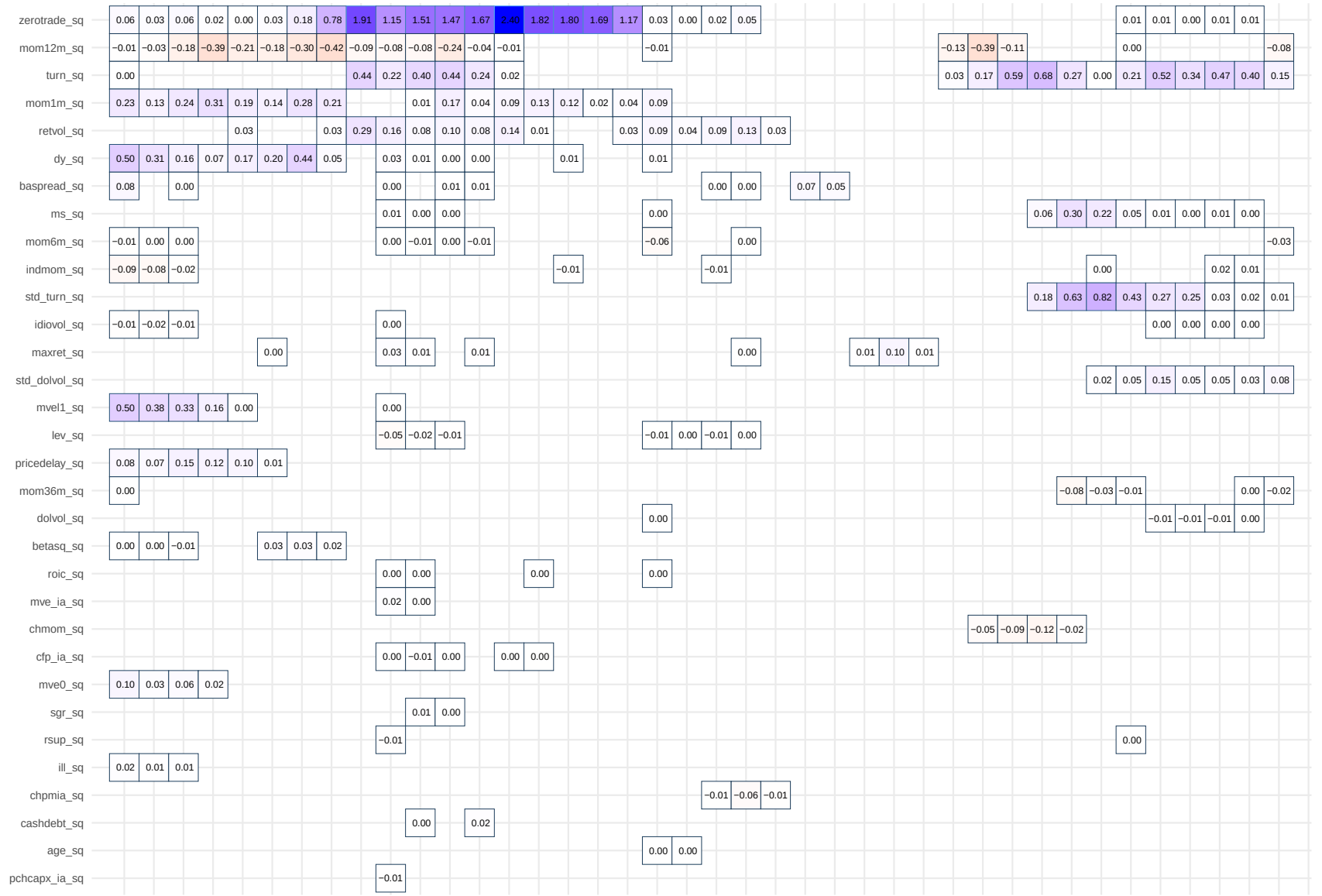


Figure 14: **Elastic net coefficients: interactions**

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

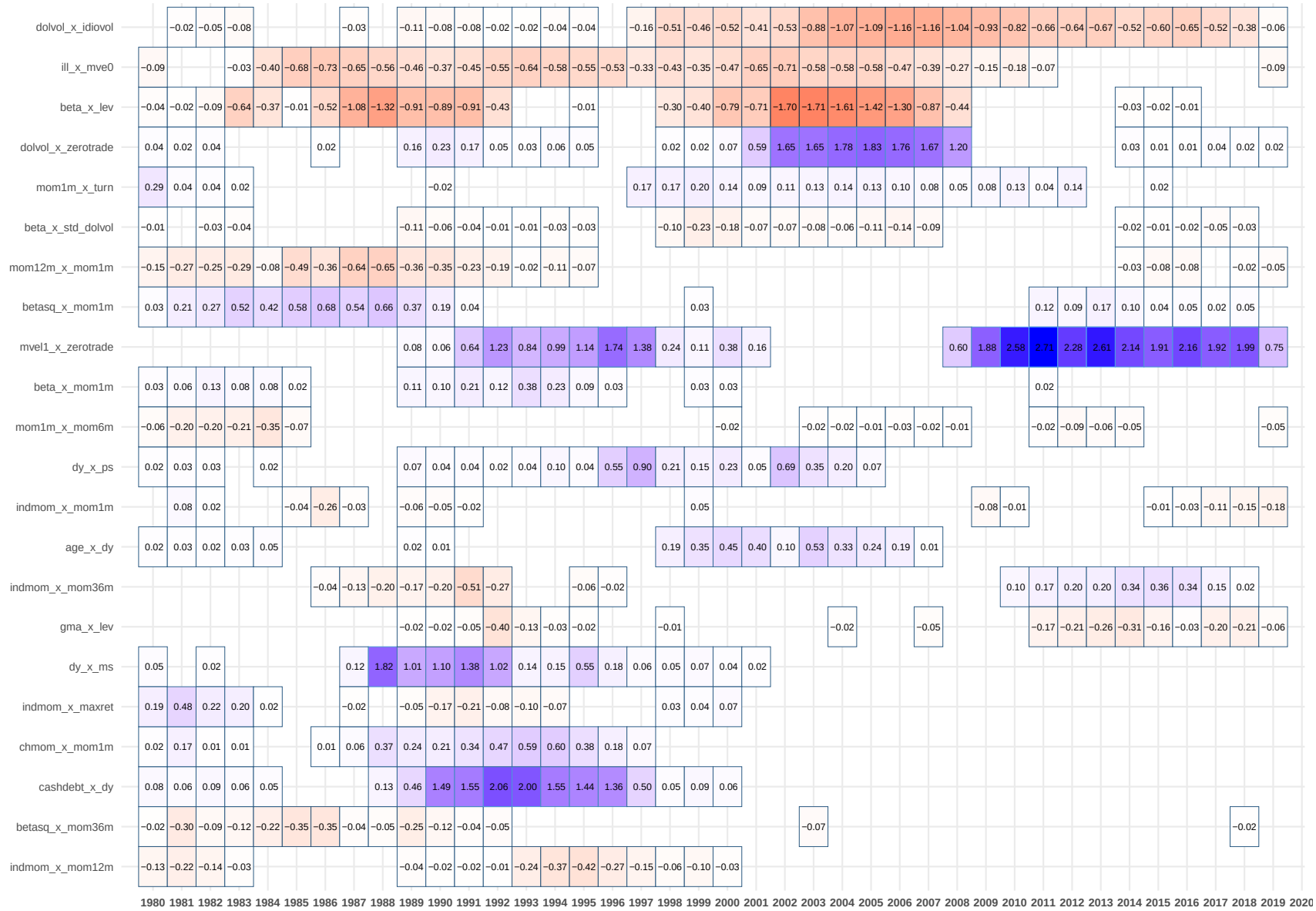


Figure 15: $L2$**-boosting coefficients: first-order effects**

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

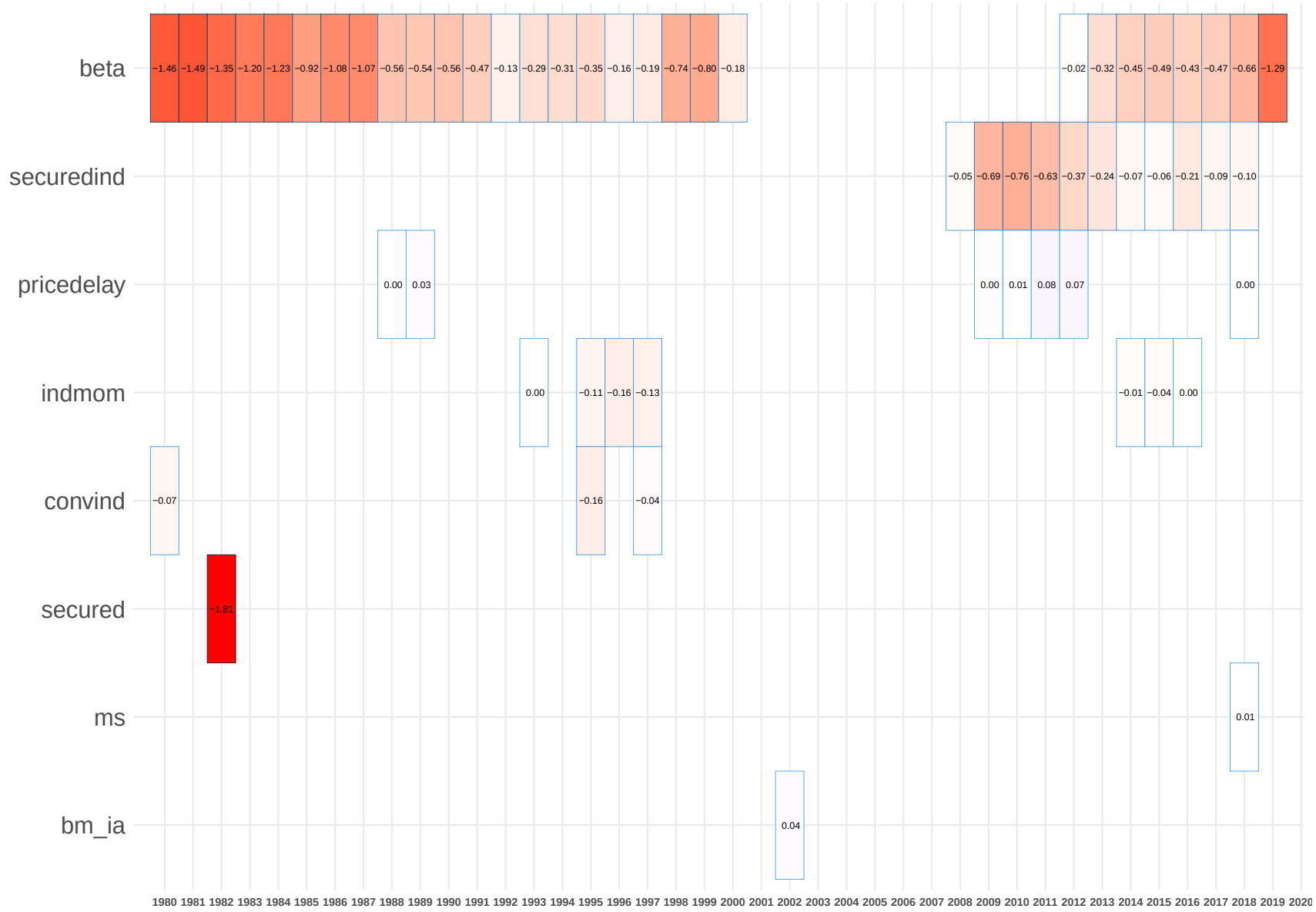


Figure 16: $L2$-**boosting coefficients: second-order effects**

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

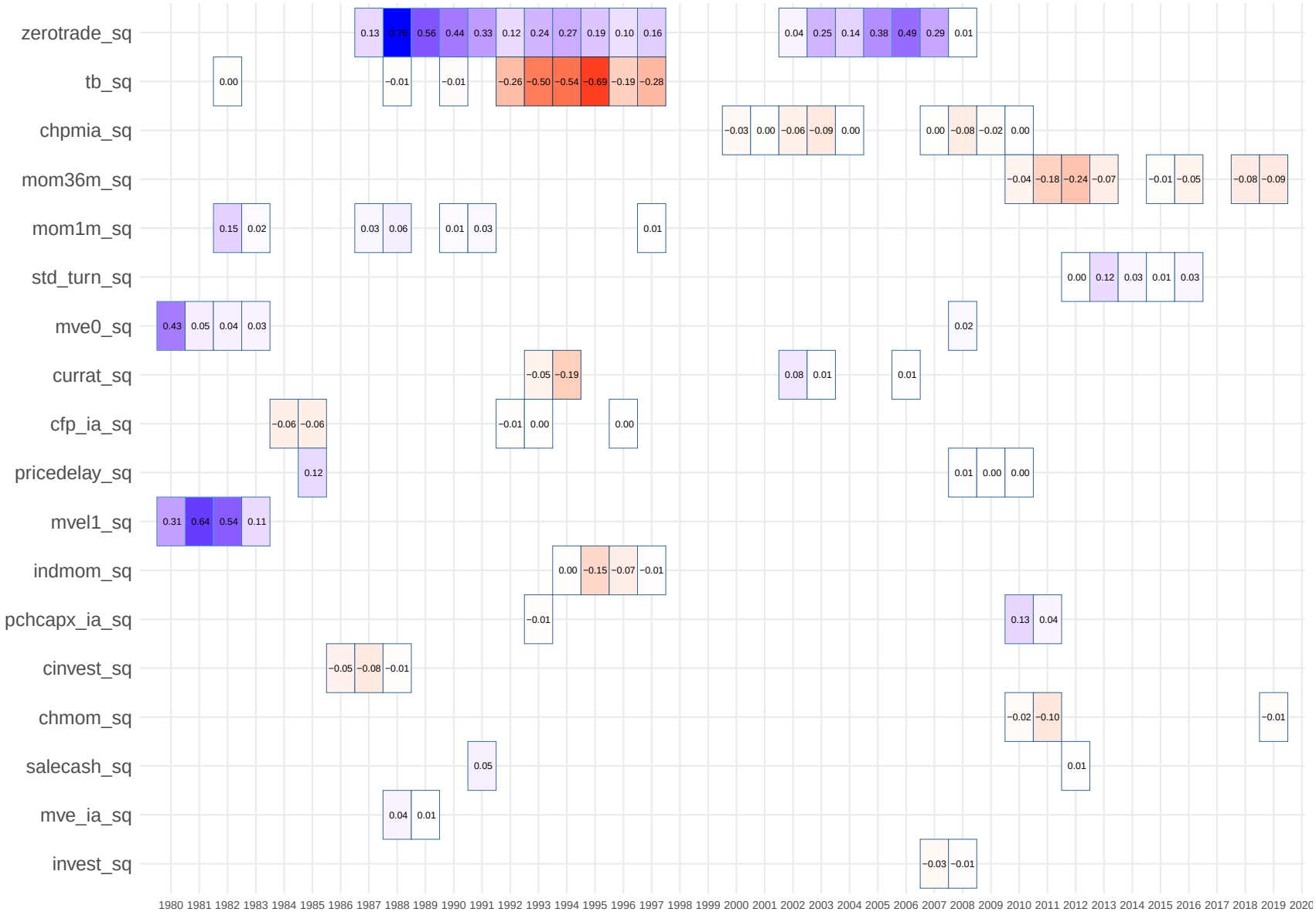
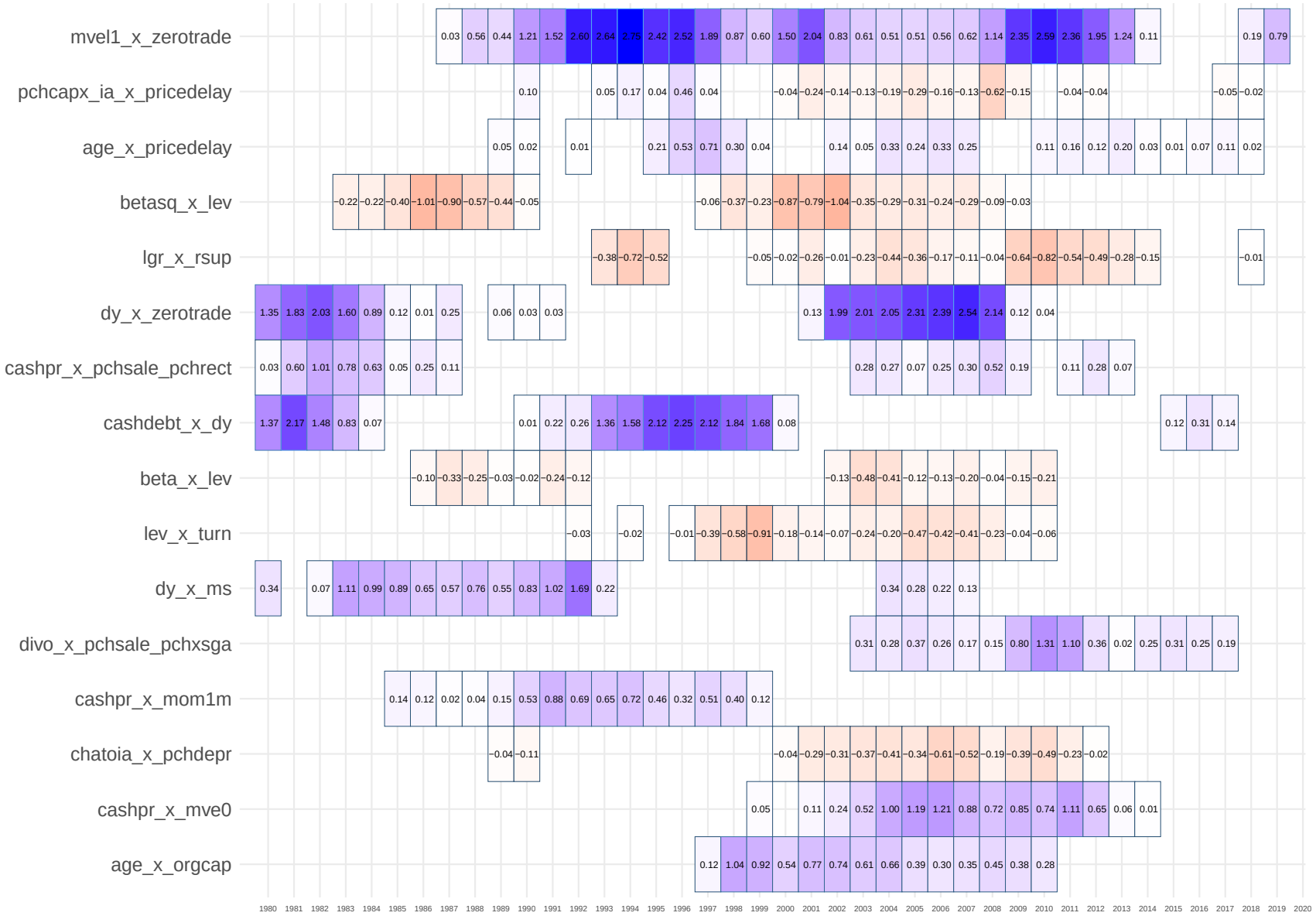


Figure 17: $L2$**-boosting coefficients: interactions**

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 $\hat{\theta}^{[m]}_{k^*}$. 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:

$$\hat{\theta}_k^{[m]} = -\frac{\hat{\sigma}_{b_{m-1}c_k}}{\hat{\sigma}^2_{c_k}}$$

    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 $\hat{\theta}^{[m]}_{k^*}$
    Update $\hat{\theta}_m$ using a step size ν:

$$\hat{\theta}_m = \hat{\theta}_{m-1} + \nu\hat{\theta}^{[m]}, m \geq 1$$

    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

1: Compute the product $X^T y$ to obtain a vector in $\mathbb{R}^p$
2: Compute the absolute values $\left|X^T y\right|$ to get the marginal correlations
3: Identify $\hat{S}$, the index set of the k largest entries in $\left|X^T y\right|$
4: Extract the submatrix $X_{\hat{S}}$ from X using the indices in $\hat{S}$
5: 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 $\hat{\beta}_{\hat{S}}$