What Explains Latent Demand?

Patrick Brock* & Max Geilen†

January 31, 2025

Abstract

We revisit the demand-based asset pricing framework introduced by Koijen & Yogo (2019) to investigate the role of latent demand. One of their key findings is that demand drives stock return volatility, but it remains unexplained which components drive demand. To tackle this issue, we extend their model by incorporating 60 additional characteristics identified in the empirical asset pricing literature. However, these additional factors do not systematically reduce the importance of latent demand. Our findings suggest that the characteristics-based demand equation does not adequately capture investor demand and cannot account for stock return volatility. We identify significant limitations in using 13F portfolio data for demand-based asset pricing frameworks due to investor heterogeneity. We suggest future research directions, emphasizing improved clustering methods, sentiment-based factors, and endogenous stock supply.

*Goethe University Frankfurt
†Goethe University Frankfurt

Contents

List of Figures

List of Tables

1 Introduction

Since Koijen & Yogo (2019) [hereafter KY19], demand-based asset pricing [hereafter DBAP] has gained popularity. The core idea of DBAP is rooted in the understanding that financial markets price equities based on price signals, a concept established at least since Fama (1970). However, the exact determination of this fair value remained largely unexplored until KY19. Their main argument is that the fair value of a stock is updated by the interplay of supply and demand subject to market clearing. Figure 1 illustrates the mechanism: investors receive signals and trade financial assets based on those signals. The new fair value is the price that satisfies the market clearing condition under the updated demand.

The main contribution of KY19 is that they are able to derive a characteristic-based demand function for stocks that directly links an investor's portfolio holdings to stock fundamentals. To estimate investor demand, they use institutional portfolio holdings data while accounting for selection biases and endogeneity issues. Finally, they use their estimated characteristics-based demand function to construct an asset demand system (market clearing mechanism) that computes a stock's new market clearing price (updated fair value) in the presence of new signals from which they can obtain counterfactual returns.

One of the main results of KY19 is presented in their Table 3, where they are able to decompose the volatility of stock returns into individual supply and demand components. Under exogenous stock supply, they are able to show that demand drives about 90% of stock return volatility, but they must attribute almost all of the impact of demand to unexplained demand shocks, which they refer to as *latent demand*. Latent demand includes all factors that are omitted from the model. This may include private information that leads an investor to desire the idiosyncratic risk of an asset or characteristics not included in the model. In short, KY19 found that demand is the key driver of stock return volatility, but they cannot assess how demand systematically affects stock return volatility.

Despite numerous contributions to DBAP (see below), this result of KY19 has not prompted further research addressing the variance decomposition of stock returns. In light of e.g. the excess volatility puzzle (Shiller 1981), KY19 have shown that DBAP can provide fundamental insights, so we address this gap in the literature.

Therefore, we adopt the framework of KY19, who used a limited number of 5 characteristics, and extend it by the large number of anomalies identified by the empirical asset pricing literature (Chen & Zimmermann 2022) to mitigate the role of latent demand and to extract systematic demand effects. We use all obviously non-endogenous anomalies consisting of all return invariant factors. In total, we include 60 additional factors. To handle this high dimensionality, we have built an extensive and flexible Python library, since the original work of KY19 runs on Stata, which is not only too slow to handle this

extension in a reasonable amount of time, but also in our attempts using KY19's code, the estimation failed too many times to be classified as reliable. For example, for the key parts of the GMM Estimation and the Variance Decomposition, we have roughly written 2,000 lines of code with an extensive documentation. In addition, because Python is a non-proprietary programming language, anyone can use our code, greatly reducing the barrier to entry for researchers wanting to engage in DBAP. Moreover, we provide a comprehensive and detailed description of KY19's framework based on their code, which also lowers the barrier to entry because their textual explanations are more expository of the DBAP implementation.

Our main finding is that additional factors *do not* mitigate the role of latent demand. Various variable selection routines, including LASSO, backward selection, and tree-based methods, fail to identify any relevant characteristic other than market capitalization, i.e. the price of the stock. The implications are twofold: either investors fail to account for anomalies, or the KY19 framework is deficient. We provide ample evidence in support of the second. First, we are able to show that the characteristics-based demand equation fails to explain investor demand, often yielding worse predictions than the simple average. While the goodness of fit can be significantly improved by using penalized nonlinear least squares estimation with fivefold cross-validation, this approach also fails to mitigate the impact of latent demand. Thus, capturing investor demand with the characteristics-based demand equation cannot explain the volatility of stock returns.

Furthermore, we show that the commonly used 13F portfolio data are suboptimal for estimating investor demand. This is because most investors must be grouped together, as they hold too few assets individually to reasonably estimate their demand, which enforces grouped investors demand function to be homogeneous. However, we show that enforcing homogeneous demand across grouped is neither a valid imposition nor a reasonable approximation as grouped investors clearly do not exhibit homogeneous preferences. Consequently, further research is necessary to identify additional attributes, beyond investor type and assets under management, to categorize investors who are ex-ante likely to demonstrate similar preferences. Additionally, we demonstrate that GMM Estimation of the characteristics-based demand equation yields biased partial effects. This is a significant limitation, as the key object of study in the literature of DBAP is the price elasticity of demand.

Overall, our results imply that KY19's framework should neither be used for policy applications, since changes in demand *cannot* be reliably predicted by the characteristics-based demand equation, nor should it be used to estimate price elasticities of demand.

However, we *do not* doubt that investor demand plays a decisive role in determining the volatility of stock returns or other aspects in asset pricing, since asset demand is a crucial factor in financial markets. We suggest that future research in DBAP could benefit from the identification of appropriate factors, such as sentiment, and the development of a functional form of investor demand that can reliably predict demand. This endeavor is

indeed challenging, as the functional form must meet monotonicity conditions to ensure the existence of a market clearing price. Furthermore, we advocate for the implementation of more sophisticated clustering methodologies in comparison to the rudimentary categorization of investors based on their type and assets under management. The formation of more homogeneous groups will serve to enhance the validity of any estimation framework employed. More sophisticated methods can also be used to predict the "zeros" which are stocks managers actively didn't buy even though they could have in order to tackle the selection bias and also mitigate the extensive margin of latent demand. Finally, the household sector, which is quite substantial, is merely a residual entity. Superior data could illuminate this sector more thoroughly (see the comments in KY19 regarding this matter) and further improve DBAP.

Our work is related to several others. Instead of estimating the level of portfolio weights, van der Beck (2022*b*) suggests estimating the change in portfolio weights, i.e. the trades. He finds that this mitigates the bias from unobservable portfolio tilts, which we and KY19 refer to as the selection bias. This bias arises because the 13F portfolio holdings data only captures realized investor demand and not stocks that investors actively chose not to buy even though they could have. However, estimating trades does not change the need to group investors. van der Beck (2022*b*) doesn't run into this problem because it focuses exclusively on mutual funds, which hold a large number of stocks and thus produce enough trades for a valid estimation procedure. Nevertheless, van der Beck (2022*b*) reports an R^2 of about 1% for trade prediction and concludes that "the stock-specific variables used in KY (2019) do not appear to be the primary characteristics that investors respond to when making their quarterly trades," a finding that is entirely consistent with our results. In other words, due to this limited explanatory power, estimating trades instead of levels does not mitigate the role of latent demand. In addition, estimating trades will make it more difficult to find market clearing prices, since during the root-finding process hypothetical trades could be computed such that level holdings are negative, which would violate any short selling conditions.

van der Beck (2022*a*) is a nice application of van der Beck (2022*b*). There, he uses his method of estimating trades to extract the price elasticity of demand for mutual funds. He estimates that withdrawing \$1 from the market portfolio and investing it in an ESG portfolio increases the total market capitalization of green stocks by about \$0.3 to \$0.5. This elasticity is used to compute counterfactual returns, which is the same logic underlying the variance decomposition in KY19, the tool used to assess the systematic impact of demand on the volatility of stock returns. The author finds by simulation that if the flows into ESG funds were instead reinvested in an aggregate mutual fund portfolio, the ESG fund return would be 200 basis points lower. In other words, by artificially capping flows into ESG funds, the price elasticity of demand is used to calculate a hypothetical return. However, no market clearing mechanism is used and supply is also exogenous[1],

[1]This is not explicitly stated, but the paper uses proposition 1.6 in van der Beck (2022*b*), where the proof sets supply to be exogenous (see text below equation (1.22)).

so this is a less sophisticated approach than that used by KY19. Moreover, the role of latent demand is simply not part of the research question.

Although we, like many others working on DBAP, take supply to be exogenous, Sammon & Shim (2024) provide evidence that supply is endogenous. They find that firms are the primary sellers of stock when index funds are net buyers. Specifically, when passive investors demand 1 percentage point of a stock's outstanding shares, firms take the other side of the trade and adjust the supply of shares by 0.64 percentage points on average. Thus, firms have an upward sloping demand curve and are also identified as the single largest group in accommodating index fund demand. They also point out that only 20% of this increase in supply is due to buybacks or seasoned equity offerings, and the remaining 80% is due to equity compensation. In other words, employees receive shares and sell them in the market when prices rise because an inelastic index fund buys shares. However, there is no built-in mechanism for repurchase, so if prices fall due to index funds selling shares, employees will not repurchase those shares. The authors make explicit reference to DBAP and the potential benefits of endogenizing supply in light of their results. For example, with endogenous supply, manifested by an upward sloping curve, prices rise less than with exogenous demand.

Fuchs, Fukuda & Neuhann (2023) raise fundamental concerns about the theoretical foundations of the asset demand system described in KY19. They refer to demand driven by non-pecuniary motives such as ESG scores as taste, and show that a no-arbitrage condition for assets doesn't hold in a neoclassical framework with tastes. To illustrate, consider two assets, A and B, with identical certain payoffs but different taste characteristics, and two investors, 1 and 2, with heterogeneous tastes. Because of his tastes, investor 1 derives a higher utility from holding asset A than from holding asset B, and because of investor 2's taste, he derives a higher utility from holding asset B than from holding asset A. Both would be willing to pay a positive price for a portfolio that is long in their preferred asset and short in their non-preferred asset. Two agents willing to pay positive prices for the opposite position violate the no-arbitrage condition.
Furthermore, they show that the low demand elasticities found in demand-system asset pricing models don't contradict the high demand elasticities of neoclassical models, but are to be expected. We expect high demand elasticities for two assets with very similar payoffs when, ceteris paribus, the price of one of the assets changes. In the data, however, we observe the demand elasticity only after endogenous price adjustments of assets that are substitutes. The authors give an example of an exogenous shock to demand for Apple stock that would also endogenously change the price of Microsoft stock. Therefore, we would expect to observe only a small demand elasticity since we can't observe it ceteris paribus.

Koijen, Richmond & Yogo (2023) adapt KY19 and develop a framework for quantifying the impact of market developments and regulatory changes on asset prices, price informativeness, and wealth distribution. They follow KY19 closely, but differ in the grouping of investors. First, they restrict the demand coefficients to be constant within each year,

with only an investor-specific intercept varying from quarter to quarter. KY19 uses quarterly estimation, but by extending the cross-section to the whole year, an investor can hold more shares and thus his demand is easier to estimate. For the remaining investors that need to be grouped, they also use a different procedure. As in KY19, they assume homogeneous behavior of each investor in the group and estimate the coefficients, but then they estimate each investor individually using the group estimates as a shrinking target (see their equation (16) & (17)). Unfortunately, they do not report any R^2, so we cannot assess whether this leads to a better performance of the characteristics-based demand equation.

Other applications of DBAP include the following: Haddad, Huebner & Loualiche (2025) use an asset demand system to estimate how the rise of passive investing changed the elasticity of demand and found that it made it 11% more inelastic. Li (2022) uses fund flows to estimate price changes in the Fama-French size and value factors. Bahaj, Czech, Ding & Reis (2023) use transaction-level data on the universe of traded UK inflation swaps to characterize who buys and sells inflation risk, when, and with what price elasticity. Han, Roussanov & Ruan (2021) find that underperforming mutual funds increase their market risk by buying stocks with higher beta. They use an asset demand system to compute counterfactual prices without the increased demand from mutual funds. The counterfactual returns don't show the low-beta anomaly.

2 Model

This section presents the framework introduced by KY19. To elucidate the intricacies, a slight extension of the notation is necessary, which not only facilitates the adoption of the framework for the reader but also enables the exposition of the problems regarding the estimation. As in KY19, the supply of stocks is treated as exogenous and is not modeled.

2.1 Demand

Definitions. We augment the notation of KY19 because the notation they use not permit a detailed disposition of the variance decomposition, the central metric we analyze.

- Time $t \in \mathbb{T} = \{0, 1, 2, \dots, \}$ is discrete and denotes quarterly increments.
- The set of investors at each time t is denoted by $\mathbb{I}_t$. Each manager has a unique identifier that is consistent over time. We denote the number of managers at time t by $I_t = |\mathbb{I}_t|$.
- Each investor $i \in \mathbb{I}_t$ has assets under management (AUM) denoted by $A_{i,t}$.
- The set of inside assets at time t consists of stocks traded on the NYSE, AMEX and NASDAQ. This set is denoted by $\mathbb{S}_t$.
  - Furthermore, the outside asset, i.e. any portfolio of assets not in $\mathbb{S}_t$, is indexed by $n = 0$. The outside asset acts as a residual because investors are not required

to invest all of their AUM in $\mathbb{S}_t$.

- The subset of inside assets an investor invested in, i.e. his portfolio, is denoted by $\mathcal{P}_{i,t} \subseteq \mathbb{S}_t$.[2]
- Lastly, characteristics of each asset $n \in \mathbb{S}_t$ are denoted by a vector $\boldsymbol{x}_t(n)$ of dimension $K \times 1$ where K denotes the number of characteristics which includes a constant, the log of market equity and other firm characteristics.

Characteristic-Based Demand Equation. KY19 derive the demand function at time t of an investor $i \in \mathbb{I}_t$ resulting from a portfolio choice problem with *no short-selling*. In doing so, they assume that investors can have heterogeneous expectations of factor returns, which allows for heterogeneous optimal demands $\delta_{i,t}(n)$ between investors. This demand demand function, coined the *characteristics-based demand equation* by KY19, is given in the following

$$\delta_{i,t}(n) \equiv \frac{\mathrm{w}_{i,t}(n)}{\mathrm{w}_{i,t}(0)} = exp\left\{\boldsymbol{x}_t'(n)\boldsymbol{\beta}_{i,t}\right\}\varepsilon_{i,t}(n) \qquad \forall t \in \mathbb{T}, i \in \mathbb{I}_t, \forall n \in \mathcal{P}_{i,t}, \tag{2.1}$$

where w are portfolio weights and $\boldsymbol{\beta}_{i,t}$ is the to be estimated $K \times 1$ vector of loadings of demand onto stock characteristics $\boldsymbol{x}_t(n)$. The scalar $\varepsilon_{i,t}(n) \geq 0$ is an error component[3], referred to by KY19 as *latent demand*, which captures the effect of all unobserved characteristics and/or private information of an investor not included in the model.

Once $\boldsymbol{\beta}_{i,t}$ in (2.1) has been estimated, the portfolio weight of the outside asset is computed by the following

$$\sum_{n\in\mathcal{P}_{i,t}} \mathrm{w}_{i,t}(n) + \mathrm{w}_{i,t}(0) = 1 \quad \Leftrightarrow \quad \Big[\sum_{n\in\mathcal{P}_{i,t}} \delta_{i,t}(n) + 1\Big]\mathrm{w}_{i,t}(0) = 1 \quad \Leftrightarrow \quad \mathrm{w}_{i,t}(0) = \frac{1}{1 + \sum\limits_{n\in\mathcal{P}_{i,t}} \delta_{i,t}(n)}.$$

Then, the portfolio weights for the remaining assets in the investment universe are given by the product of relative demand and the portfolio weight of the outside asset so that

$$\mathrm{w}_{i,t}(n) = \frac{\delta_{i,t}(n)}{1 + \sum_{k\in\mathcal{P}_{i,t}} \delta_{i,t}(k)} \qquad \forall n \in \mathcal{P}_{i,t}. \tag{2.2}$$

Equation (2.2) also clarifies that the portfolio weight depends on the characteristics of all other stocks in $\mathcal{P}_{i,t}$, i.e. there are cross-dependencies.

Notice that the investors' portfolio weights explicitly depend on the stocks' prices as the relative portfolio weight (2.1) depends on stock characteristics $\boldsymbol{x}_t$ which contain log market equity (log supply plus log price).

[2] $\mathcal{P}_{i,t}$ makes the notation more substantial than in KY19 as will become apparent in Section 4.1.
[3] The inequality must hold as demand can never be negative due to the no short-selling restriction.

2.2 Market Clearing

Using exogenous supply and the portfolio weights given by (2.2), KY19 aim to find a market clearing price such that stock supply is equal to stock demand. This is what they refer to as the *Demand System Approach to Asset Pricing*. A stock's price is the key variable as all policy experiments are evaluated by finding the new hypothetical market clearing price used to compute hypothetical returns.

Definitions. To establish market clearing, we firstly define the following objects.

- Let $\mathcal{P}_t := \bigcup_{i \in \mathbb{I}_t} \mathcal{P}_{i,t}$ denote at time t the set of inside stocks for which at least one investor has a strictly positive portfolio weight[4].
- Let $N_t := |\mathcal{P}_t|$ denote the number of stocks in $\mathcal{P}_t$.
- Investor demand is captured by the vector $\boldsymbol{q}_{i,t}$ of dimension $N_t \times 1$. This vector contains the number of shares demanded for each stock in $\mathcal{P}_t$ from investor $i \in \mathbb{I}_t$. Furthermore, $q_{i,t}(n) = 0 \; \forall n \notin \mathcal{P}_{i,t}$.[5]
- An investor's vector of portfolio weights is given by $\boldsymbol{w}_{i,t}$ which is of dimension $N_t \times 1$. Likewise, $w_{i,t}(n) = 0 \; \forall n \in \mathcal{P}_t \setminus \mathcal{P}_{i,t}$ and $w_{i,t}(n)$ is given by (2.2) $\forall n \in \mathcal{P}_{i,t}$.
- *Log* stock prices are captured by the vector $\boldsymbol{p}_t$ of dimension $N_t \times 1$.
- *Log* stock supply is captured by the vector $\boldsymbol{s}_t$ of dimension $N_t \times 1$ which contains the *log* shares outstanding.
  - As mentioned beofre, $\boldsymbol{p}_t + \boldsymbol{s}_t = log(\boldsymbol{ME}_t)$ where ME is the abbreviation for market equity, which we se to be the first element of $\boldsymbol{x}_t$.
- Let $\boldsymbol{a}_t$ of dimension $I_t \times 1$ contain the AUM of each investor.
- Stock characteristics are captured by matrix $\boldsymbol{X}_t$ of dimension $(K-1) \times N_t$ which contains all characteristics (rows) apart from log market equity for each stock (columns).
  - Log market equity is omitted as it is endogenously determined by $\boldsymbol{s}_t + \boldsymbol{p}_t$.
- Analogously, let matrix $\boldsymbol{B}_t$ of dimension $I_t \times K$ store all the individual $\boldsymbol{\beta}_{i,t}$ in (2.1) which remain to be appropriately estimated (see Section 4). For now, they are taken as given.
- Lastly, let matrix $\boldsymbol{E}_t$ of dimension $I_t \times N_t$ store of each investor i (rows) his latent demand for every asset $n \in \mathcal{P}_t$ (columns) where $\varepsilon_{i,t}(n) = 0 \; \forall n \in \mathcal{P}_t \setminus \mathcal{P}_{i,t}$ and else-wise it is given by (2.1), i.e. $\varepsilon_{i,t}(n) = \delta_{i,t}(n) exp\left\{-\boldsymbol{x}_t'(n)\boldsymbol{\beta}_{i,t}\right\} \; \forall n \in \mathcal{P}_{i,t}$.

[4] If no investor holds a stock even though it is in $\mathbb{S}_t$, then demand is zero by default and finding a market clearing price such that supply equals demand is not applicable.

[5] Notice that $n \in \mathcal{P}_t$ are not ascending counting numbers so that in theory a mapping is required that maps n to the index position in the vector $\boldsymbol{q}_{i,t}$, which are counting numbers. In practice, we simply sort the vector in ascending order by the stock identifier *permno*, an integer number. We won't formalize such a mapping as it will simply make the notation more cumbersome without adding any value. This remark applies to all further presented vectors.

Aggregate Demand. As investors cannot short-sell, an individual investor's demand $\boldsymbol{q}_{i,t}$ is determined by

$$\boldsymbol{q}_{i,t} \odot exp\{\boldsymbol{p}_t\} = A_{i,t}\boldsymbol{w}_{i,t}, \tag{2.3}$$

where "$\odot$" denotes the Hadamard product (element-wise multiplication). In other words, the number of shares held times the price per share is equal to the fraction of AUM the investor invests into the stock. Then, aggregate demand for all stocks in $\mathcal{P}_t$ is simply given by

$$\boldsymbol{Q}_t = \sum_{i \in \mathbb{I}_t} A_{i,t}\boldsymbol{w}_{i,t}.$$

Notice that each element in $\boldsymbol{Q}_t$ is strictly positive because we only consider $n \in \mathcal{P}_t \subseteq \mathbb{S}_t$.

Market Clearing. As aggregate supply is simply the vector of market equity of the stocks, the market clearing price $\boldsymbol{p}_t^*$ must satisfy

$$log\Big(\underbrace{\boldsymbol{Q}_t(\boldsymbol{p}_t^*|\boldsymbol{s}_t, \boldsymbol{X}_t, \boldsymbol{a}_t, \boldsymbol{B}_t, \boldsymbol{E}_t)}_{\text{Aggregate Demand}}\Big) \quad - \quad \underbrace{(\boldsymbol{s}_t + \boldsymbol{p}_t^*)}_{\text{Log Aggregate Supply}} = 0, \tag{2.4}$$

where it was made explicit in the notation that $\boldsymbol{Q}_t$ is a function of the running variable $\boldsymbol{p}_t$ given shares outstanding $\boldsymbol{s}_t$, AUMs $\boldsymbol{a}_t$, the remaining characteristics $\boldsymbol{X}_t$, the coefficients $\boldsymbol{B}_t$ and latent demand $\boldsymbol{E}_t$, i.e. all the other information required apart from the price to compute the portfolio weights in (2.2).

KY19 prove in their Appendix that (2.4) always admits a market clearing price $\boldsymbol{p}_t^*$ if the demand curve is downward sloping for which they derive the restriction (2.7) given further below.

Household Sector. The household sector is vital for the construction of the asset-demand system, i.e. for (2.4). For the observed data at time t, a household sector is constructed which holds the remaining shares outstanding of stocks in $\mathcal{P}_t$. This residual sector is necessary because the investors in $\mathbb{I}_t$ naturally needn't buy the entire market share of the stocks in $\mathcal{P}_t$. Thus, the construction of the household sector ensures that the market is always cleared for the observed data. For this household data a characteristics-based demand equation is also estimated. The term *Household* used by KY19 might be somewhat ambiguous, as it includes all investors not listed in the 13F Files, rather than solely retail investors. Figure 3 illustrates the ratio of the aggregated AUM of 13F investors to the AUM of the household sector, highlighting that the household sector is substantial. This significance is not only evident in the aggregate but also at the individual stock level, where the household sector holds a notable share in the majority of stocks. This observation is particularly relevant in light of Section 5.1, which indicates that the preferences of the household sector are complex and challenging to determine precisely.

Demand Elasticity. Considering (2.3) in logs for which the elements are strictly positive,

the price elasticity of demand is given by (see Appendix A for a proof)

$$-\frac{\partial log\big(\boldsymbol{q}_{i,t}\big)}{\partial \boldsymbol{p}_t} = I_{|\mathcal{P}_{i,t}|} - diag(\boldsymbol{w}_{i,t})^{-1}\beta_{1,i,t}\Big[diag(\boldsymbol{w}_{i,t}) - \boldsymbol{w}_{i,t}\boldsymbol{w}_{i,t}'\Big] \qquad \forall n, j \in \mathcal{P}_{i,t}, \tag{2.5}$$

where $\beta_{1,i,t}$ is the coefficient of market equity and $I_{|\mathcal{P}_{i,t}|}$ is the $|\mathcal{P}_{i,t}| \times |\mathcal{P}_{i,t}|$ identity matrix. Thus, a Jacobian of dimension $|\mathcal{P}_{i,t}| \times |\mathcal{P}_{i,t}|$ captures all (cross) elasticities for stocks for which the investor has strictly positive portfolio weights. Notice that the individual elasticities are given by

$$-\frac{\partial log\big(\boldsymbol{q}_{i,t}(n)\big)}{\partial \boldsymbol{p}_t(n)} = 1 - \beta_{1,i,t}[1 - w_{i,t}(n)] \qquad \forall n \in \mathcal{P}_{i,t}. \tag{2.6}$$

KY19 prove that as long as

$$\beta_{1,i,t} < 1, \tag{2.7}$$

i.e. as long as demand is downward sloping, a market clearing price that satisfies (2.4) always exists. The intuition is that if $\beta_{1,i,t} > 1$, then an increase in the price of a stock will increase demand. Since supply is exogenously fixed, higher demand will automatically lead to further price increases, which in turn increase demand so that a price spiral towards infinity is set off.

2.3 Variance Decomposition

Upon establishing the *asset demand system* in (2.4), it is now possible to decompose the variance of annual stock returns into demand and supply components. Annual stock returns from June are used because most companies' fiscal years end in December, so the characteristics are typically updated in June of the following year. We describe in detail how the variance decomposition works, in particular the separation of latent demand effects, based on the Stata code provided by KY19, which is considerably more nuanced than their stylized explanation on p. 1505.

In essence, as the world is slowly "updated" from t to $t+4$, the variance of stock returns is decomposed by determining new intermediate market clearing prices that must adhere to market clearing (2.4) at each update step. These new intermediate market clearing prices allow the calculation of intermediate returns, which are then used to decompose the variance of the final annual return. The decomposition is formalized below.

Firstly, from the definition of returns from t to next year $t+4$, i.e. $R_{t,t+4} = \frac{P_{t,t+4}+D_{t,t+4}}{P_t}$, we obtain $r_{t,t+4} = log(P_{t,t+4}[1 + D_{t,t+4}/P_{t,t+4}]) - p_t$ by taking the log. This annual log return for every asset is captured in a vector

$$\boldsymbol{r}_{t,t+4} = \boldsymbol{p}_{t,t+4} - \boldsymbol{p}_t + \boldsymbol{v}_{t,t+4},$$

with $\boldsymbol{v}_{t,t+4} \equiv log(1 + exp\{\boldsymbol{d}_{t,t+4} - \boldsymbol{p}_{t,t+4}\})$ being the dividend return taken as given, i.e. it

is not influenced by changes in demand.

Next, as in KY19, market clearing condition (2.4) is summarized by a function

$$\boldsymbol{g}(\boldsymbol{s}, \boldsymbol{X}, \boldsymbol{a}, \boldsymbol{B}, \boldsymbol{E})$$

outputting the market clearing price $\boldsymbol{p}^*$ solving (2.4) given $\boldsymbol{s}, \boldsymbol{X}, \boldsymbol{a}, \boldsymbol{B}$ & $\boldsymbol{E}$. Hence, function $\boldsymbol{g}$ solves a root-finding problem. We solve this high-dimensional root finding problems with *SciPy*'s Newton-Krylov solver which is 12 times faster than the algorithm provided by KY19. The L2 norm of the large vector with length over 3000 is typically in the order of 10^{-14}. Our replication package contains all numerical errors at all steps for all scenarios run.

The variance decomposition works as follows.

1. Take all investors in $\mathbb{I}_t$ and only consider assets in $\mathcal{P}_t$.

   (a) The initial market clearing price $\boldsymbol{p}_t^* = \boldsymbol{g}(\boldsymbol{s}_t, \boldsymbol{X}_t, \boldsymbol{a}_t, \boldsymbol{B}_t, \boldsymbol{E}_t)$ is given from the data.

   (b) Update shares outstanding and obtain the intermediary market clearing price

   $$\boldsymbol{p}_{t|\boldsymbol{s}} = \boldsymbol{g}(\widetilde{\boldsymbol{s}}_{t+4}, \boldsymbol{X}_t, \boldsymbol{a}_t, \boldsymbol{B}_t, \boldsymbol{E}_t),$$

   where $\widetilde{\boldsymbol{s}}_{t+4}$ contains next year's number of shares outstanding for all stocks that continue to be listed. For delisted stocks, the contemporaneous shares outstanding are used[6]. Thus, $\widetilde{\boldsymbol{s}}_{t+4}$ is not the same object as the observed shares outstanding vector $\boldsymbol{s}_{t+4}$ from the data because $\widetilde{\boldsymbol{s}}_{t+4}$ contains stocks in $\mathcal{P}_t$ and $\boldsymbol{s}_{t+4}$ contains stocks in $\mathcal{P}_{t+4}$ which generally never completely overlap.

   (c) Repeat the previous step iteratively where values are updated when updates exist and else-wise the contemporaneous value is kept[7] to obtain intermediary prices

   $$\begin{aligned}
   \boldsymbol{p}_{t|\boldsymbol{s},\boldsymbol{X}} &= \boldsymbol{g}(\widetilde{\boldsymbol{s}}_{t+4}, \widetilde{\boldsymbol{X}}_t, \widetilde{\boldsymbol{a}}_{t+4}, \boldsymbol{B}_t, \boldsymbol{E}_t) \\
   \boldsymbol{p}_{t|\boldsymbol{s},\boldsymbol{X},\boldsymbol{a}} &= \boldsymbol{g}(\widetilde{\boldsymbol{s}}_{t+4}, \widetilde{\boldsymbol{X}}_{t+4}, \widetilde{\boldsymbol{a}}_{t+4}, \boldsymbol{B}_t, \boldsymbol{E}_t) \\
   \boldsymbol{p}_{t|\boldsymbol{s},\boldsymbol{X},\boldsymbol{a},\boldsymbol{B}} &= \boldsymbol{g}(\widetilde{\boldsymbol{s}}_{t+4}, \widetilde{\boldsymbol{X}}_{t+4}, \widetilde{\boldsymbol{a}}_{t+4}, \widetilde{\boldsymbol{B}}_{t+4}, \boldsymbol{E}_t).
   \end{aligned}$$

2. Take all investors in $\mathbb{I}_{t+4}$ and only consider assets in $\mathcal{P}_{t+4}$.

   (a) Downgrade latent demand back to time period t if the investor previously existed in t *and* additionally held the stock in t, else-wise keep contemporaneous latent demand as the downgrade doesn't exist and compute the intermediary

[6] Considering Figure 7, this is far from uncommon.

[7] For example, if investors cease to exist, their contemporaneous AUM is kept.

price

$$\boldsymbol{p}_{t+4|\boldsymbol{E}} = \boldsymbol{g}(\boldsymbol{s}_{t+4}, \boldsymbol{a}_{t+4}, \boldsymbol{X}_{t+4}, \boldsymbol{B}_{t+4}, \widetilde{\boldsymbol{E}}_t).$$

This step here is the extensive margin which KY19 label as a ceteris paribus change of the investment universe. Notice that the objects $\boldsymbol{p}_{t+4|\boldsymbol{E}}, \boldsymbol{s}_{t+4}, \boldsymbol{a}_{t+4}, \boldsymbol{X}_{t+4}, \boldsymbol{B}_{t+4}$ & $\widetilde{\boldsymbol{E}}_t$ are of different dimension than in step 1 as both the number of investors and the stocks have been updated.

(b) Update latent demand to $\boldsymbol{E}_{t+4}$ so that the market clearing price

$$\boldsymbol{p}^*_{t+4} = \boldsymbol{g}(\boldsymbol{s}_{t+4}, \boldsymbol{a}_{t+4}, \boldsymbol{X}_{t+4}, \boldsymbol{B}_{t+4}, \boldsymbol{E}_{t+4})$$

is the observed price from the data.

3. Consider only the intersection of stocks in $\mathcal{P}_t \cap \mathcal{P}_{t+4}$.

(a) Compute the following intermediate returns

$$\begin{aligned}
\Delta\boldsymbol{p}_t(\boldsymbol{s}) &= \boldsymbol{p}_{t|\boldsymbol{s}} - \boldsymbol{p}^*_t \\
\Delta\boldsymbol{p}_t(\boldsymbol{X}) &= \boldsymbol{p}_{t|\boldsymbol{s},\boldsymbol{X}} - \boldsymbol{p}_{t|\boldsymbol{s}} \\
\Delta\boldsymbol{p}_t(\boldsymbol{a}) &= \boldsymbol{p}_{t|\boldsymbol{s},\boldsymbol{X},\boldsymbol{a}} - \boldsymbol{p}_{t|\boldsymbol{s},\boldsymbol{X}} \\
\Delta\boldsymbol{p}_t(\boldsymbol{B}) &= \boldsymbol{p}_{t|\boldsymbol{s},\boldsymbol{X},\boldsymbol{a},\boldsymbol{B}} - \boldsymbol{p}_{t|\boldsymbol{s},\boldsymbol{X},\boldsymbol{a}} \\
\Delta\boldsymbol{p}_t(\boldsymbol{E}_1) &= \boldsymbol{p}_{t+4|\boldsymbol{E}} - \boldsymbol{p}_{t|\boldsymbol{s},\boldsymbol{X},\boldsymbol{a},\boldsymbol{B}} \\
\Delta\boldsymbol{p}_t(\boldsymbol{E}_2) &= \boldsymbol{p}^*_{t+4} - \boldsymbol{p}_{t+4|\boldsymbol{E}}.
\end{aligned}$$

Notice that these vectors have a different dimension than in the previous step 2 and 1.

(b) Using the intermediary prices, the annual log return for every asset is given by

$$\boldsymbol{r}_{t,t+4} = \Delta\boldsymbol{p}_t(\boldsymbol{s}) + \Delta\boldsymbol{p}_t(\boldsymbol{X}) + \Delta\boldsymbol{p}_t(\boldsymbol{a}) + \Delta\boldsymbol{p}_t(\boldsymbol{B}) + \Delta\boldsymbol{p}_t(\boldsymbol{E}_1) + \Delta\boldsymbol{p}_t(\boldsymbol{E}_2) + \boldsymbol{v}_{t,t+4}.$$

4. Compute the right hand side of $\boldsymbol{r}_{t,t+4}$ for the entire time series, stack them all and denote this by omitting the time subscript. Then, writing $V(\boldsymbol{r}) = Cov(\boldsymbol{r}, \boldsymbol{r})$ and inserting the above, the variance decomposition due to the linearity of the covariance is simply given by computing the following empirical covariances of the hypothetical intermediary returns with the observed log returns

$$\begin{aligned}
1 = &\frac{Cov\big(\boldsymbol{r}, \Delta\boldsymbol{p}(\boldsymbol{s})\big)}{V(\boldsymbol{r})} + \frac{Cov\big(\boldsymbol{r}, \Delta\boldsymbol{p}(\boldsymbol{X})\big)}{V(\boldsymbol{r})} + \frac{Cov\big(\boldsymbol{r}, \boldsymbol{v}\big)}{V(\boldsymbol{r})} + \frac{Cov\big(\boldsymbol{r}, \Delta\boldsymbol{p}(\boldsymbol{a})\big)}{V(\boldsymbol{r})} \\
&+ \frac{Cov\big(\boldsymbol{r}, \Delta\boldsymbol{p}(\boldsymbol{B})\big)}{V(\boldsymbol{r})} + \frac{Cov\big(\boldsymbol{r}, \Delta\boldsymbol{p}(\boldsymbol{E_1})\big)}{V(\boldsymbol{r})} + \frac{Cov\big(\boldsymbol{r}, \Delta\boldsymbol{p}(\boldsymbol{E_2})\big)}{V(\boldsymbol{r})}
\end{aligned}$$

such that the sum is always equal to one so that each summand can be interpreted

as a contribution (positive or negative) to the overall stock return variance. Notice that these fractions can be computed from an OLS regression which has the further benefit of obtaining standard errors and allowing for the implementation of time fixed effects.

Remarks. KY19 does not address this issue, but it is unclear how to deal with stocks or investors that are not present in two subsequent time periods. It is not clear whether stocks that only exist in period t should be included and simply not updated or whether they should be excluded from the beginning. The same applies to managers that cease to exist in the next period. Are they still driving some of the stock return variance, and if so, to what extent? The extensive margin is a natural error component because there is no modeling of the introduction of new investors or the exit of existing ones, and likewise for assets. In an additional analysis, we considered each subsequent time period t and $t+4$, and focused on investors who were present in both periods. For these investors, we only included the stocks that they held in both periods. However, this approach broke down the variance decomposition (see Section 6.3).

3 Data

3.1 Characteristics Data

In order to estimate the $\boldsymbol{\beta}_{i,t}$ in (2.1) data on stock characteristics $\boldsymbol{x}_t(n)$ and portfolio holdings $\delta_{i,t}$ are required.

We use data for stock prices, dividends, returns, and shares outstanding from the Center for Research in Security Prices (CRSP) monthly stock database. We only use ordinary common shares which are listed on the New York Stock Exchange, the American Stock Exchange or Nasdaq and are US domestic and no real estate investment trust. KY19 use log book equity, profitability, investment, dividends to book equity and market beta as explanatory variables. We follow their methodology and construct the stock characteristic except market beta with Compustat data. Market beta is the coefficient of a regression of monthly excess returns, over the 1-month Treasury-bill rate, onto excess market returns using a 60-month moving window (with at least 24 months of non-missing returns).
All observations with missing values for these characteristics are removed from the sample. Any asset, which is held by an investor, but not part of the sample is considered part of the outside asset.
We get data on additional firm characteristics from open source asset pricing (Chen & Zimmermann (2022)). Due to the endogeneity of latent demand and asset prices, we drop all characteristics, which are constructed using return or price data of the previous quarter in any form. Out of 212 available characteristics which are classified as return predictors, we end up with 124 characteristics which do not use returns of the previous quarter. Out of these 124 characteristics we drop any that have a coverage of less than 70% in any quarter of our sample. After applying these filters we end up with a sample

of 59 additional characteristics, bringing the total up to 65 characteristics including the 6 original characteristics from KY19. Table 7 provides an overview of all characteristics used. We impute missing characteristic values with their cross-sectional mean value following Chen & Zimmermann (2022), Kozak et al. (2020), Gu et al. (2020) or Green et al. (2017). We lead all characteristic based on accounting data by six months to ensure that they have been publicly available at the point in time we use them. We lead all other characteristics by one month. Finally we winsorize the highest and lowest 2.5% of values of all non-dummy characteristic. Figure 2 displays the correlation across the entire sample period for the features. It is clear that the features add new information whereas the baseline characteristics used by KY19 (top left corner) are moderately correlated with market equity. As market equity is the most important variable, since it contains the price, it is vital that additional characteristics don't pick up this information.

Portfolio holdings data is sourced from the SEC's quarterly Form 13F filings, which are mandatory quarterly filings existing since 1980 for institutional investors managing over $100 million. These filings, typically submitted 45 days after each quarter's end, list the number of shares the investors hold of a stock. For more information on 13F filings, refer to KY19.

To access this data, KY19 utilize the proprietary Thomson Reuters Institutional (13F) Holdings S34 database, which is not accessible to us. Instead, we use the freely available Form 13F filings data from the SEC's EDGAR platform. This alternative, however, limits our earliest reporting date to the year 2000, as opposed to the 1980 starting date in the Thomson Reuters data. The 1980 to 2000 period is exclusive to Thomson Reuters because the SEC contracted Thomson Reuters to process these filings which until the year 2000 were submitted on paper or microfiche (Wharton Research Data Services (2017)).

Although the period 2000 to 2012 is freely downloadable from EDGAR, it is provided in a non-standardized text format. Fortunately, Backus, Conlon & Sinkinson (2021) have scraped and cleaned this data, making it freely available (see their Section II). We downloaded this data from Michael Sinkinson's webpage. The data from 2013 onward is provided in a standardized table format by the SEC which we directly downloaded from EDGAR. Backus et al. (2021), KY19, Ben-David et al. (2021) and Wharton Research Data Services (2017) document issues with the Thomson Reuters data so that the SEC data is of higher quality[8].

It is important to acknowledge the broader limitations associated with Form 13F filings data. For example, the SEC on its webpage states that "*because the data is derived from information provided by the individual filers, we cannot guarantee the accuracy of the data sets*"[9]. Anderson & Brockman (2018) provide empirical evidence of discrepancies of Form 13F filings data. Nevertheless, no other extensive data for a wide range US institutional

[8]KY19 also account for issues with the Thomson Reuters data and from 2013 onward substitute them with the freely available data provided on EDGAR.

[9]https://www.sec.gov/data-research/sec-markets-data/form-13f-data-sets
[Last Accessed: November 30th 2024]

investors is available since investors do not willingly disclose their holdings. The following table provides an example snippet of the dataset. Figure 4, 5, 6 & 7 display expositions of the dataset.

Table 1: Example Snippet of the Data.

Time	Manager ID	Stock ID	AUM	Portfolio Weight	...
⋮	⋮	⋮	⋮	⋮	⋮
2009q2	110	10001	100	0.05	...
2009q2	110	10084	100	0.01	...
⋮	⋮	⋮	⋮	⋮	⋮
2009q2	285	10012	800	0.2	...
2009q2	285	10084	800	0.08	...
⋮	⋮	⋮	⋮	⋮	⋮

4 Estimation

4.1 Investment Universe / Selection Bias

A challenge in estimating characteristics-based demand equation (2.1) is that 13F filings only provide data on *realized* demand which is captured by $\mathcal{P}_{i,t}$. However, investors do not only decide to purchase certain stocks, they also decide to *not* purchase certain stocks even though they could have. There are two primary reasons for this.

1. Investment Mandate Restrictions: An investor may be prohibited from holding certain assets due to their investment mandate. For instance, a fund specializing in the communication sector might not be permitted to invest in healthcare stocks.

2. Preference-Based Decisions: The investor might find other stocks more appealing, resulting in zero demand for certain assets.

Both scenarios result in an asset not being held, but they convey different information about the investor's preferences. The first case is independent of the investor's preferences because they are simply not allowed to hold those assets, so these assets can be excluded from the estimation. However, excluding assets from the second case would introduce a bias because it would systematically exclude assets the investor dislikes. In short, the stocks investors purchased and didn't purchase convey information, but we only have information on the purchased stocks. In other words, the data has a *selection bias*.

To distinguish between these two types of assets, KY19 cite evidence that investors are subject to (non-observable) exogenous investment mandates. They use these investment mandates to construct an investment universe $\mathcal{N}_{i,t}$ such that it includes all stocks an investor held within the current quarter or the previous 11 quarters. They motivate this choice empirically (see their Table 1), but it remains an ad-hoc choice. There is no logit or probit model that determines the likelihood of an asset to be in the investment universe.

KY19 in section B1 state that there are problems measuring the investment universe so that the configuration of $\mathcal{N}_{i,t}$ can be seen as a rough proxy, but it is unclear if it is a true reflection of the investment universe. Since investment mandates are not public, improving the measurement of the investment universe is far from a trivial task.

As a result, (2.1) can be written more clearly. Since $\mathcal{P}_{i,t} \subseteq \mathcal{N}_{i,t}$ is the set of stocks in the investor's portfolio with strictly positive weights obtained from the 13F filings (realized demand), we can re-write (2.1) more explicitly as

$$\frac{\mathrm{w}_{i,t}(n)}{\mathrm{w}_{i,t}(0)} = \begin{cases} exp\left\{\boldsymbol{x}_t'(n)\boldsymbol{\beta}_{i,t}\right\}\varepsilon_{i,t}(n), & \forall n \in \mathcal{P}_{i,t} \\ 0, & \forall n \in \mathcal{N}_{i,t} \setminus \mathcal{P}_{i,t} \end{cases} \tag{4.1}$$

so that $\mathcal{N}_{i,t} \setminus \mathcal{P}_{i,t}$ is the set of stocks in the investment universe with *zero* portfolio weight (unrealized demand). These weights are referred to as zero holdings or the "zeros" by KY19. These "zeros" are a majority portion of the dataset as Figure 8 shows.

As the exponential function is strictly positive regardless of $\boldsymbol{\beta}_{i,t}$, this entails that

$$\varepsilon_{i,t}(n) = 0 \qquad \forall n \in \mathcal{N}_{i,t} \setminus \mathcal{P}_{i,t}. \tag{4.2}$$

Estimating (4.1) poses challenges because the dependent variable can be zero, making it impossible to apply a *log* transformation. In fact, KY19 show in their Appendix that neglecting the zeros leads to biased estimates.

To estimate (4.1) unbiasedly, we must ensure that

$$\begin{aligned} \mathbb{E}[\varepsilon_{i,t}(n)|\boldsymbol{x}_t(n)] &= \underbrace{\mathbb{P}\big(\varepsilon_{i,t}(n) > 0|\boldsymbol{x}_t(n)\big)}_{\in(0,1)} \cdot \underbrace{\mathbb{E}\big[\varepsilon_{i,t}(n)|\varepsilon_{i,t}(n) > 0, \boldsymbol{x}_t(n)\big]}_{=1} \\ &\quad + \underbrace{\mathbb{P}\big(\varepsilon_{i,t}(n) = 0|\boldsymbol{x}_t(n)\big)}_{\in(0,1)} \cdot \underbrace{\mathbb{E}\big[\varepsilon_{i,t}(n)|\varepsilon_{i,t}(n) = 0, \boldsymbol{x}_t(n)\big]}_{=0} \\ &= c_{i,t} \in (0,1) \qquad \forall t \in \mathbb{T}, i \in \mathbb{I}_t, n \in \mathcal{N}_{i,t}. \end{aligned} \tag{4.3}$$

where $c_{i,t}$ is a constant strictly inside the unit interval. KY19 use this same decomposition (4.3) in their footnote 10, but *incorrectly* set $\mathbb{E}[\varepsilon_{i,t}(n)|\boldsymbol{x}(n)] = 1$ for all time periods and for all investors[10] which implies that $\mathbb{E}\big[\varepsilon_{i,t}(n)|\varepsilon_{i,t}(n) > 0, \boldsymbol{x}(n)\big] > 1$, resulting in a bias (see Figure 9).

However, in practice the most relevant aspect is that $\mathbb{E}[\varepsilon_{i,t}(n)|\boldsymbol{x}(n)]$ is equal to a constant as this is absorbed into the constant of $\boldsymbol{x}'\boldsymbol{\beta}$. Nonetheless, this complicated structure of the error no longer leads to consistent estimates of the partial effects, see Section 6.2.

To enhance comparability, we follow KY19 and use

$$\mathbb{E}[\varepsilon_{i,t}(n)|\boldsymbol{x}(n)] = 1 \qquad \forall t \in \mathbb{T}, i \in \mathbb{I}_t, n \in \mathcal{N}_{i,t}. \tag{4.4}$$

[10] Only in the first period where there are no "zeros" by construction of $\mathcal{N}_{i,t}$ does this actually hold.

4.2 Identification

Due to endogeneity, $\mathbb{E}\big[\varepsilon_{i,t}(n)|\boldsymbol{x}(n)\big] = 1$ will not hold. This is because the investors considered are far from atomistic so that the error ε (latent demand) and market equity are very likely correlated. In other words, demand shocks from investors will influence prices. Therefore, an instrument for market equity is used. To address this endogeneity issue, KY19 introduce an instrumental variable for log market equity which exploits the cross sectional variation in the investment universe and wealth of investors. This instrument is given by the following

$$\widehat{me}_{i,t}(n) = log\left(\sum_{j\neq i} A_{j,t}\frac{\mathbb{1}_{j,t}(n)}{1+|\mathcal{N}_{j,t}|}\right) \qquad \forall t \in \mathbb{T}, i \in \mathbb{I}_t, \forall n \in \mathcal{N}_{i,t},$$

where $\mathbb{1}_{j,t}(n)$ is the indicator function that is 1 if $n \in \mathcal{N}_{j,t}$ and 0 else. Thus, the instrumental variable represents the sum of the fractions of wealth that other investors would allocate to asset n if their portfolios were equally weighted and asset n is part of their investment universe. This instrument is exogenous and thus independent of latent demand for two reasons. Firstly, it relies on the exogenous investment universe which is predetermined (see Section 2 and Appendix E in KY19). Secondly, KY19 argue that the wealth distribution across other investors, i.e. their assets under management A, is predetermined and exogenous to contemporary demand shocks (see their Section 2 for evidence presented). Therefore, instead of (4.4), we can only impose

$$\mathbb{E}[\varepsilon_{i,t}(n)|\boldsymbol{z}(n)] = 1 \qquad \forall t \in \mathbb{T}, i \in \mathbb{I}_t, n \in \mathcal{N}_{i,t}, \tag{4.5}$$

where $\boldsymbol{z}$ is the vector of IVs. Since all other characteristics apart from market equity remain exogenous, $\boldsymbol{z}$ only replaces market equity with its instrument $\widehat{me}$ and otherwise is identical to the vector $\boldsymbol{x}$.

4.3 Pooling Investors

One significant obstacle in estimating equation (4.1) is the data itself. While large ETFs may include thousands of stocks, providing abundant data points for estimating their $\boldsymbol{\beta}_{i,t}$, most investors hold only a limited number of assets, as noted by KY19. Figure 10 illustrates this by displaying a truncated histogram of the number of individual stock holdings per investor for each year. Many investors maintain concentrated portfolios, with the median number of stocks held being approximately 70 throughout the sample. Estimating each investor's demand individually is therefore not feasible.

To address this issue, KY19 pool investors with fewer than 1,000 individual strictly positive stock holdings into *bins* based on their AUM and type. These grouped investors play a crucial role in stock price fluctuations (see their Table 4). The resulting bins have an average of approximately 2,000 positions in their combined portfolios. This binning is essential, as roughly 60% of all bins comprise of pooled investors, i.e. more than one

investor in a bin. However, there is neither a clear justification for this pooling approach nor an a priori reason to assume that investors with similar AUM and type exhibit comparable behavior. Section 6 shows that this is indeed not the case which is the cause for serious issues with the estimation.

On top of that, problems arise when more than one investor within the same bin hold the same stock. To avoid issues, an investor-specific intercept (fixed effect) is used to enhance the model fit, as illustrated in Figure 11. KY19 do not mention this in their paper, but from their code it is clear that they implement this individual specific intercept which is given by the mean log relative portfolio weight of the individual investor.

Thus, for each time point let $\mathbb{B}_t = \{B_{1,t}, \dots, B_{g,t}, \dots, B_{N_B,t}\}$ be the set of bins of investors with N_B being the number of different bins such that $\{B_{g,t}\}_{g=1}^{N_B}$ forms a partition of $\mathbb{I}_t$, i.e. $\bigcup_{g=1}^{N_B} B_{g,t} = \mathbb{I}_t$ with $B_{g_1,t} \cap B_{g_2,t} = \emptyset$ for any $g_1 \neq g_2$. Thus, $\forall i \in B_{g,t}$ we use the data points $\sum_{i \in B_{g,t}} |\mathcal{N}_{i,t}| > 1000$ to estimate $\boldsymbol{\beta}_{g,t}$ in

$$\delta_{i,t}(n) \equiv \frac{\mathrm{w}_{i,t}(n)}{\mathrm{w}_{i,t}(0)} = \begin{cases} exp\left\{\boldsymbol{x}_t'(n)\boldsymbol{\beta}_{g,t} + \widetilde{x}_t(n)\right\} \varepsilon_{i,t}(n), & \forall n \in \mathcal{P}_{i,t} \\ 0, & \forall n \in \mathcal{N}_{i,t} \setminus \mathcal{P}_{i,t} \end{cases} \tag{4.6}$$
$$\text{with } \widetilde{x}_t(n) := log\Big(\frac{1}{|\mathcal{P}_{i,t}|} \sum\nolimits_{k \in \mathcal{P}_{i,t}} \frac{\mathrm{w}_{i,t}(k)}{\mathrm{w}_{i,t}(0)}\Big)$$

where $\boldsymbol{\beta}_{g,t}$ is the demand of the investors in a bin so that each investor in the same bin has the same $\boldsymbol{\beta}_{g,t}$, i.e. $\forall i, j \in B_{g,t}\ \boldsymbol{\beta}_{i,t} = \boldsymbol{\beta}_{j,t}$. If a bin only consist of one investor, (4.6) reduces back to (4.1) where the extra term is absorbed by the constant. We follow KY19 exactly in constructing the bins.

4.4 Instrumental Variable Generalized Method of Moments

Building upon the foundations laid in Sections 4.1, 4.2 & 4.3, which culminate in equation (4.6), i.e. the basis for the estimation, and incorporating the constraint (2.7), we are now equipped to estimate coefficients $\boldsymbol{\beta}$ in the characteristics-based demand equation (2.1).

IV-GMM Setup. KY19 implement an IV-GMM estimator, but are completely silent on the implementation, which is why we explain it in detail. In doing so, we provide a comprehensive numerical solution scheme which can handle much more than the exact identified case in GMM-estimation and also, if necessary, deal with multiple domain restrictions.

To keep notation simple, we consider a fixed time point t along with a fixed bin B_g and omit the relevant subscripts. Each estimation is conducted for every bin at every time point which results in more than 18,000 individual estimations[11].

Since (4.6) is used for the estimation, it is clear that $\boldsymbol{\beta}$ cannot be estimated by OLS because a *log* transformation is not possible. In fact, KY19 show in their Appendix

[11]Notice that this makes t-tests difficult. Using a Bonferroni correction, the significance level is likely too low and will result in many false acceptances of the H_0. KY19 themselves do not report any t-statistics

that ignoring the "zeros" leads to biased estimates. This is why they resort to IV-GMM estimation. The moment condition they use is the same moment condition as in OLS, namely

$$Cov(\boldsymbol{z}, \varepsilon) = 0 \quad \Rightarrow \quad \mathbb{E}[\boldsymbol{z}\varepsilon] - \mathbb{E}[\boldsymbol{z}]\mathbb{E}[\varepsilon] = 0 \quad \overset{(4.5)}{\Rightarrow} \quad \mathbb{E}[\boldsymbol{z}(\varepsilon - 1)] = 0$$

$$\text{where } \varepsilon(n) \overset{(4.6)}{=} \begin{cases} \frac{\mathrm{w}(n)}{\mathrm{w}(0)} exp\left\{-\boldsymbol{x}'(n)\boldsymbol{\beta} - \widetilde{x}(n)\right\}, & \forall n \in \mathcal{P} \\ 0, & \forall n \in \mathcal{N} \setminus \mathcal{P} \end{cases}$$

The number of moment conditions are K (as $\boldsymbol{z}$ is of dimension $K \times 1$) so that the model is exactly identified as $\boldsymbol{\beta}$ is of dimension K.

The empirical moment condition is simply given by

$$\overline{\boldsymbol{g}}(\boldsymbol{\beta}) = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{g}_n(\boldsymbol{\beta}) \qquad \text{with} \qquad \boldsymbol{g}_n(\boldsymbol{\beta}) = \boldsymbol{z}(n)[\varepsilon(n) - 1]$$

and N being the number of observations in the bin so that n indexes an observation in the bin. Each observation is a stock holding (potentially a zero holding) of each investor in the bin.

Then, in accordance with Hansen (1982), the IV-GMM estimator satisfies the following

$$\boldsymbol{\beta}^{IV}_{GMM} = \arg\min_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}) \tag{4.7}$$

$$\text{with } Q(\boldsymbol{\beta}) = \overline{\boldsymbol{g}}(\boldsymbol{\beta})'\boldsymbol{W}(\boldsymbol{\beta}^{IV}_{GMM})\overline{\boldsymbol{g}}(\boldsymbol{\beta}) \quad \text{and} \quad \boldsymbol{W}(\boldsymbol{\beta}) = \left(\frac{1}{N}\sum_{n=1}^{N} \boldsymbol{g}_n(\boldsymbol{\beta})\boldsymbol{g}_n(\boldsymbol{\beta})'\right)^{-1},$$

where $\boldsymbol{W}(\boldsymbol{\beta}^{IV}_{GMM})$ is the optimal weighting matrix. A practical problem to determine the optimal weighting matrix is that the IV-GMM estimator requires the optimal weighting matrix, but the optimal weighting matrix in turn requires the IV-GMM estimator. To break this dependence, an iterative approach can be used to reach the fixpoint. Thus, take $\boldsymbol{W}^{(0)} = \boldsymbol{I}$ and get $\boldsymbol{\beta}^{(0)}_{IVGMM}$ from minimizing Q. Iteratively use $\boldsymbol{\beta}^{(i)}_{IVGMM}$ to compute $\boldsymbol{W}^{(i+1)} = \left(\frac{1}{N}\sum_{n=1}^{N} \boldsymbol{g}_n(\boldsymbol{\beta}^{(i)}_{IVGMM})\boldsymbol{g}_n(\boldsymbol{\beta}^{(i)}_{IVGMM})'\right)^{-1}$ and solve for $\boldsymbol{\beta}^{(i+1)}_{IVGMM}$ until the weighting matrix converges.

For an exactly identified problem as is the case here, the optimal weighting matrix is irrelevant as the objective function Q can be set to zero exactly. Therefore, an equivalent simplified criterion is to simply find $\boldsymbol{\beta}^{IV}_{GMM}$ such that

$$\overline{\boldsymbol{g}}(\boldsymbol{\beta}^{IV}_{GMM}) = 0 \tag{4.8}$$

which is a root-finding instead of a minimization problem.

Restriction. The main challenge now is to incorporate restriction (2.7). We establish a two-step procedure:

1. Assume the restriction is non-binding and simply solve root-finding problem (4.8).
    - We do this first because solving (4.8) only takes a split second.
2. If $\beta_1 \geq 1$, solve minimization problem (4.7) with an L-BFGS-B solver to incorporate the restriction.
    - As KY19, we impose $\beta_1 \leq 0.99$ because such solvers can only incorporate weak inequalities.

Solving (4.7) is actually a difficult task, which is why we will outline it in detail. Nonetheless, the task is worth achieving because being able to solve (4.7) has many advantages such as being able to impose many more restrictions if ever found to be necessary or being able to incorporate more moment conditions that would for example improve the fit of the estimation. Moreover, sometimes root-finding (4.8) fails such that having a second, more stable approach, will reduce the number of bins we have to throw out in the variance decomposition[12].

Solving Root-Finding (4.8). We solve the root-finding problem by simply using the standard Newton-Algorithm. For the initial guess we use the biased 2SLS estimates of (4.6) which discard the "zeros". We stop when the L1-error of the objective function is smaller than 10^{-15} which typically takes 5 to 7 iterations. We implement a vectorization to speed up the computation and further compute the Jacobian analytically to achieve higher accuracy and further speed gains. This is derived in Appendix A.

Solving Minimisation (4.7). Firstly, the objective function $Q(\boldsymbol{\beta})$ can be immediately implemented using the vectorization approach established for (4.8). However, actually minimizing (4.7) turns out to be a particularly challenging problem irrespective of constraint (2.7). The main reason is that the exponential function quickly explodes and any minimization routine will potentially evaluate the objective function Q at some $\boldsymbol{\beta}$ where the exponential function will lead to explosive values. This problem is only exacerbated if more explanatory variables are included because just a single observation of a characteristic with a high value can cause everything to blow up. For example, $exp(20)$ is already in the hundred millions. We can summarize our solution to the overflow problem in the following 4 necessary steps.

1. Standardize the input $\boldsymbol{x}$ of the exponential function to avoid overflow errors and sensitivities in specific directions.
    - Note that due to the linear transformation in the exponential function we can recover the original $\boldsymbol{\beta}$ to ensure that the constraint (2.7) holds. This is derived in Appendix A.3.
2. Reduce the step size of the BFGS-algorithm to further avoid overflow errors.

[12]Unfortunately, KY19 do not report how often their estimation converges, so we cannot make a comparison.

3. Implement the weighting matrix with the iterative GMM-Estimator to increase numerical stability.

    - Although theoretically not necessary, the weighting matrix optimally scales the moment conditions, placing greater emphasis on conditions with lower variance and down-weighting those with higher variance. This reduces the influence of noisier moments, leading to faster convergence to the true parameter values. Moreover, moment conditions that correlate highly with others are up-weighted as minimizing such moment conditions minimizes others along with them.

4. Apply the BFGS-algorithm iteratively starting with a high tolerance level that is reduced step-by-step to the desired tolerance level.

    - Although this itself does not necessarily prevent overflow errors as the other 3 steps, this step is the most important as it achieves convergence. The reason is that when the BFGS solver resets it will reset its initialization of the Hessian, so that the BFGS solver will not get stuck in local minima.

The restriction (2.7) is simply implemented by an *L-BFGS-B* solver. On top of that, we also compute the Jacobian of objective function Q with automatic differentiation. This is not necessary, but it does increase the accuracy and convergence speed slightly. Using this routine, for the unconstrained case the MSE is 10^{-32} and the results were identical to the ones as when solving the root-finding problem. Naturally, when solving the constrained problem, the MSE is much higher. Please check our replication package where we document all numerical errors for every bin in every setting.

Alternative way to tackle the Restriction. While it is possible to use the domain restriction $\beta_1 \equiv 1 - e^{-\widetilde{\beta}_1}$ with $\widetilde{\beta}_1 \in \mathbb{R}$ such that $\beta_1 \in (-\infty, 1)$, this substitution complicates the computation and vectorization of the Jacobian. Moreover, when minimizing the MSE it complicates the standardization of the coefficients as the bound cannot be immediately adjusted. Moreover, this transformation has led to the root-finding to fail sometimes. Since more than $18,000$ estimations each for numerous settings need to be computed, a robust procedure was preferred.

Comparison with KY19. In the provided replication code the estimation is conducted in Stata where they use four initial guesses and two different solver methods (Gauss-Newton and modified Newton-Raphson). However, we were unsuccessful in our attempts to get the Stata code to converge so that we cannot compare the runtime or numerical errors as KY19 do not report this.

4.5 Replication

Here we briefly document our replication of KY19. We present these results to demonstrate the comparability of our code and data with their findings. Figure 12 displays the average coefficients by investor type, weighted by assets under management (comparable

to Figure 3 in KY19). A higher coefficient on market equity indicates a lower demand elasticity.

Our results confirm the findings of KY19 that banks, mutual funds, and pension funds exhibit a lower demand elasticity. This aligns with the hypothesis that these institutions are more constrained due to their large size and the benchmarking of their investment mandates. Additionally, we replicate their observation that households tend to tilt their portfolios towards high-dividend and low-profitability stocks. Among institutional investors, banks demonstrate the highest demand for stocks paying high dividends.

Regarding the key metric analyzed, Table 2 reports the replication of the variance decomposition (compare with Table 3 in KY19). As can be seen, the results match well, thereby reiterating justification for comparison regarding our extension.

Table 2: Variance Decomposition Baseline

	% of Variance
Supply:	
Shares Outstanding	3.62
	(0.73)
Stock Characteristics	5.94
	(0.44)
Dividend Yield	0.22
	(0.02)
Demand:	
AUM	2.35
	(0.16)
Coefficients	5.49
	(0.44)
Latent demand: extensive margin	22.86
	(0.56)
Latent demand: intensive margin	59.52
	(0.79)
Observations	63,521

Note: The cross-sectional variance of annual stock returns is decomposed into supply- and demand-side effects. Heteroskedasticity-robust standard errors are reported in parentheses. Time fixed-effects are included. Sample period is 2002-2022.

5 Variable Selection

Latent demand is the primary driver of cross-sectional variance in stock returns. However, KY19 employ only six characteristics to explain portfolio holdings. In light of the 'Factor Zoo' (Cochrane 2011), which has emerged in the empirical asset pricing literature, this number is relatively small. On top of that, including more characteristics allows for different investors to prioritize a different subset of characteristics. Given the diverse nature

of investors and their varying trading motives, investment horizons and risk appetites, it seems intuitive that not all investors pay attention to the same characteristics. For instance, an index fund's demand may be solely determined by a firm's market capitalization to replicate an index, while a hedge fund might prioritize earnings surprises over firm size. Therefore, incorporating a broader set of characteristics should, in theory, help reduce the latent demand component in the variance decomposition.

To select a subset of features, we employ various feature selection processes in order to see which characteristics investors prioritize. We firstly employ a standard LASSO where we run an OLS and also the IV2SLS Estimator to account for the endogeneity of market equity. On top of that, we further compute an adaptive LASSO since this features the oracle property (Zou 2006). For robustness we also ran a backward selection procedure with IV2SLS and also GMM. For GMM we refrained from an IV approach as this severely reduced statistical power[13]. Lastly, we use a tree based method.

However, no feature selection approach yields any meaningful insight into investors' behavior apart from market equity being the sole important variable. Details are presented below.

5.1 LASSO

We attempted to implement a LASSO and an Elastic Net GMM Estimator as proposed by Caner (2009) and Caner & Zhang (2014), but given the previously documented challenges in estimating the characteristics-based demand equation we were unable to resolve numerical challenges for these estimators.

Therefore, we use a standard LASSO estimator, thereby ignoring the "zeros" and taking the log in (2.1) in order to obtain a linear model. The estimation is conducted for each bin in each cross section separately. All characteristics are normalized to have unit variance and mean zero. We tune the hyper-parameter using 5-fold cross validation.

The main result of the *LASSO* estimation is displayed in Figure 13. This Figure shows the frequency with which the six baseline characteristics from KY19 are selected compared to the additional characteristics, based on a pooled calculation. Log market equity emerges as the most significant variable, which is expected since it is the only one to directly incorporate price information, else-wise the other characteristics wouldn't be exogenous. Moreover, many additional characteristics are at least equally important as the baseline ones. Nevertheless, apart from log market equity, no clear pattern emerges regarding which characteristics investors prioritize. Interestingly, households appear to care about numerous characteristics, despite being the least informed investors. This counterintuitive result arises because the household sector is a residual category ensuring market clearing in the data (compare Section 2.2). Thus, it does not only contain retail investors as its name suggest, but also all other non-retail residual investors not captured in the 13F

[13]In the code the results for IV can be computed by changing one macro.

filings.

Figure 14 shows the results using value-weighted averages (weighted by AUM) instead of pooled calculations. The insights remain consistent: beyond market equity, no characteristic stands out across investor types. To illustrate this, Figure 15 presents LASSO results for AQR Capital Management. Here, market equity, book equity, profit, investment, and operating leverage are the most important variables, while no consistent pattern emerges for the others. Results of the LASSO that used the instrument of market equity are omitted, as they convey the same conclusions[14].

Given that the LASSO estimation selected an excessive number of characteristics, we proceeded with an adaptive LASSO. Leng et al. (2006) show that the shrinkage parameter in LASSO is too low when tuning it to fit optimal prediction accuracy so that indeed LASSO often selects too many variables. Zou (2006) shows that LASSO performs correct model selection only under certain conditions and suggests the adaptive LASSO as an improvement.

The adaptive Lasso is a two step procedure which penalizes coefficients differently based on their importance. Individual penalty strengths solve LASSO‘s problem of a too low penalty. We implement the adaptive LASSO with the LARS estimator using OLS estimates as weights (Zou (2006), Equation (4)) for each bin in each cross-section separately. Alternative weights, such as Ridge or LASSO estimates, do not change the results and are therefore not portrayed. Figure 16 displays the results. The adaptive LASSO identifies market equity as the sole relevant variable. Value-weighted and IV results are excluded, as they lead to identical conclusions. Figure 17 illustrates the adaptive LASSO results for AQR Capital Management, reinforcing that only market equity is significant. Additionally, the adaptive LASSO occasionally deems all variables irrelevant, making a simple average the best predictor. This issue is discussed further in Section 6.

5.2 Backward Selection

Figure 19 shows results from the backward selection procedure using the IV2SLS estimator, while Figure 18 presents results from the GMM backward selection procedure for which the test statistics were computed using Newey & McFadden (1994) Theorem 3.4 (alternative source: Greene (2019) Theorem 13.2). These results also lack conclusive patterns like LASSO since they select too many variables. Given the approximately 18,000 bins tested multiple times, the significance level of $\alpha = 5\%$ leads to numerous false rejections of the null hypothesis so that by design too many variables are chosen. However, adjusting the significance level via Bonferroni correction to approximate an overall $\alpha = 5\%$ results in the same conclusions as the adaptive LASSO. Furthermore, using the IV estimator in the GMM backward selection procedure greatly reduces statistical power, so that market equity is basically never selected which is evidently not a valid result.

[14]These plots are included in the replication package.

5.3 Results

Our results show that incorporating additional characteristics and performing variable selection isn't straightforward. Variable sets chosen using LASSO include too many characteristics. The adaptive LASSO leads to an extremely sparse variable set which basically reduces to market equity. Thus, the overall explanatory power of the considered characteristics for portfolio holdings is very limited, leading to cases where it is best not to select any characteristic. Therefore, model selection in a demand based asset pricing framework remains challenging and requires further efforts to deliver convincing results. Possible further steps are to diversify the set of characteristics. The exclusion of any price related factors narrows down the characteristic set significantly. Additional instrumental variables such as the momentum instrument proposed by van der Beck (2022*b*) or sentiment could be valid alternatives. Another possible way of incorporating more information from factors to the model is through machine learning methods such as gradient boosting regressions or neural networks. Gu et al. (2020) identify these as best performing models in modeling expected returns.

While machine learning methods offer valuable insights, they may not always ensure that the parameter set is constrained in a way that guarantees market clearing (as discussed in Section 2.2). Specifically, the characteristics-based demand equation (2.1) is not be the most suitable framework for modeling investor demand, and the 13F data has limitations when it comes to predicting investor demand accurately.

Table 3 shows the results of variance decomposition for the different variable selection procedures. None of the LASSO selection methods attenuate the influence of latent demand. Interestingly, the adaptive LASSO, which basically selects only market equity as the relevant variable, reverses the importance of the error component, so that the impact of the extensive margin is greatly reduced.

While the backward selection procedures do mitigate the influence of latent demand, this is entirely due to the coefficients, which now account for a substantial portion of the variance. This is not surprising, since backward selection not only introduces significant variation in the coefficients, but also typically selects at least 20 coefficients per bin. This effect is most pronounced when considering the selection of all characteristics at all time points, with coefficients explaining nearly 45% of the total variance in stock returns - a clear indication of overfitting. In sum, no subset of characteristics is identified that satisfactorily mitigates the role of latent demand.

What also stands out is the role of stock characteristics in mitigating stock return variance. This may seem surprising, but it is not, as Table 4 shows. When updating each stock characteristic separately in KY19's framework, it is clear that even in the baseline, some stock characteristics mitigate stock return variance. This was not seen because these effects were all aggregated together. Since the different variable selection procedures select different characteristics, the negative contributions may outweigh each other so that the stock characteristics mitigate demand.

Table 3: Variance Decomposition Variable Selection

	Baseline	LASSO	LASSO IV	Adapt. LASSO	Back. Sel. IV2SLS	Back. Sel. GMM	All
Supply:							
Shares Outstanding	3.62	2.65	3.8	2.04	3.92	3.16	3.95
	(0.73)	(0.76)	(0.72)	(0.85)	(0.72)	(0.78)	(0.7)
Stock Characteristics	5.94	-7.28	-3.58	6.25	-3.94	-3.53	-2.14
	(0.44)	(3.4)	(3.65)	(1.83)	(2.72)	(3.55)	(2.94)
Dividend Yield	0.22	0.22	0.22	0.22	0.22	0.22	0.22
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Demand:							
AUM	2.35	2.91	2.3	3.86	2.27	2.74	1.81
	(0.16)	(0.21)	(0.2)	(0.24)	(0.19)	(0.21)	(0.19)
Coefficients	5.49	23.02	15.0	1.86	22.29	21.63	44.95
	(0.44)	(3.52)	(3.83)	(0.87)	(3.2)	(3.95)	(9.34)
Latent demand: extensive margin	22.86	23.54	26.49	49.72	18.82	18.21	-4.04
	(0.56)	(0.92)	(1.39)	(1.11)	(1.82)	(1.89)	(8.89)
Latent demand: intensive margin	59.52	54.94	55.76	36.06	56.41	57.57	55.25
	(0.79)	(0.92)	(0.88)	(1.94)	(0.87)	(1.03)	(0.88)
Observations	63,521	63,521	63,521	63,521	63,521	63,521	63521

Note: The cross-sectional variance of annual stock returns is decomposed into supply- and demand-side effects. Heteroskedasticity-robust standard errors are reported in parentheses. Time fixed-effects are included. Sample period is 2002-2022.

Table 4: Variance Decomposition Baseline Iterative

	% of Variance
Supply:	
Shares Outstanding	3.62
	(0.73)
Book Equity	4.24
	(0.3)
Profit	-5.66
	(0.46)
Investment	4.37
	(0.39)
Dividends to BE	-2.45
	(0.3)
Market Beta	-0.82
	(0.16)
Dividend Yield	0.22
	(0.02)
Demand:	
Assets under Management	2.51
	(0.16)
Coefficients	5.5
	(0.44)
Latent demand: extensive margin	28.95
	(0.7)
Latent demand: intensive margin	59.52
	(0.79)
Observations	63,521

5.4 Gradient Boosting Regression

KY19 choose variables which have been shown to predict stock returns to explain investor holdings. So far we extended this set of variables to increase the explanatory power of our model. In the next step we alter the methodology to see which variables are the most important to explain investor demand.

Gu et al. (2020) test a variety of machine learning techniques to predict returns. They show that tree based methods and neural networks work well to predict returns, because they can include information from nonlinear predictor interactions. Given that a typical investor wants to tilt his portfolio towards stocks with higher expected returns we use a tree based method, namely a gradient boosting regression, to explain investors demand. A gradient boosting regressions builds a tree by repeatedly splitting the data into bins based on the regressor variables until a certain depth is reached. The average value of the dependent variable in each bin serves as the prediction for all data points within that bin. The next tree is then "grown" using the residuals of the predictions of the previous tree as dependent variable and this prediction is added to those from the previous trees. While each tree itself is a weak predictor adding many together leads to a strong predictor. It is

a nonparametric estimator which doesn't need any prior assumptions on the shape of the relationship between demand and the regressors. While this allows us to not rely on the exponential function to model the characteristics based demand equation it comes with a drawback. We can't put a restriction on the relationship between price and demand and therefore not restrict the price elasticity of demand to be negative. But we need the price elasticity to be negative to guarantee that there is a unique market clearing price (see section 2.2). Therefore we focus on the gradient boosting regression in the context of variable selection. We fit a gradient boosting regression for each investor or bin of investors in each time period. Then we estimate the variable importance of each predictor and compare the results to our previous selection exercises.

In gradient boosting regression, variable importance is typically calculated based on the contribution of each feature to the model's predictive power. This is measured by aggregating the total reduction in mean squared error achieved by splits involving a given feature across all the trees in the ensemble. Each tree in the gradient boosting process is built to minimize the residual error from the previous iterations, and features that create splits with larger reductions in the loss are considered more important. The importance values are then normalized to provide relative measures of contribution. This helps to identify which features are the most meaningful for making predictions.

To estimate the models we need to tune several parameters. We tune them separately for each investor and time period. We use a a randomized grid search with 5 fold cross-validation to find the best parameters[15].
20 shows the average importance of the variables pooled over time for each investor type. The most important variable to explain investor demand is log market equity followed by log book equity. For pension funds and households log market equity accounts for 80% of the mean squared error reduction. It is the most important variable for all investor types. Log book equity is the second most important variable but following with a large margin. The third most important variable for all investors except pension funds and households is zerotradedays. It is not a size related variable but refers to the liquidity of a stock. This confirms the findings of the previous variable selection approaches, where variables relating to the size of the firm where chosen most often.

6 Model Performance

The preceding section demonstrated that using features from the "Factor Zoo" that are evidently not endogenous does not sufficiently mitigate the role of latent demand. This raises the question: why does the "Factor Zoo" fail to provide explanatory power?

This section attributes the problem to both the 13F portfolio holdings data and the characteristics-based demand equation (2.1). First, using R^2, we show that (2.1) simply

[15]For tuning we decide to use the following parameter distribution for the randomized grid search: learning rate log-uniform distributed between (0.001,1), number of tress log-uniform distributed between (10,1000), tree depth uniform distributed between (1,10),

does not fit the data presented to it. The R^2 is the perfect metric because it measures how much of the variance can be captured by the fitted values, which is obviously crucial to mitigate the role of latent demand in the variance decomposition when computing hypothetical stock returns. In addition, we show that (2.1) leads to biased partial effects. Remedies such as using penalized nonlinear least squares estimation or holding the investment universe constant to force out a source of latent demand do not solve the problem either.

These shortcomings are serious because there is no direct remedy to improve the fit of the model, since it has already been established in Figure 10 that most investors hold only concentrated portfolios. As a result, even advanced and data-intensive estimation methods from the machine learning literature are unlikely to effectively account for the role of latent demand. Even without this problem, machine learning frameworks, especially nonparametric ones, will find it very difficult to enforce the downward sloping demand restriction (see (2.6) & (2.7)). Without this restriction, no market clearing price can be guaranteed. As a result, neither counterfactual policy experiments, such as assessing the effects of quantitative easing, nor out-of-sample predictions of investor demand are possible.

For this Section, we will focus on the KY19 baseline characteristics as the variable selection didn't result in any meaningful improvement and would only worsen an in-sample overfitting problem.

Firstly, Figure 21 displays the in-sample R^2 for the *unrestricted* OLS estimates, which disregard the "zeros". We report the unrestricted estimates solely for OLS. By completely disregarding the "zeros" and reporting the unrestricted estimates, the OLS estimates provide the best in-sample fit, representing the best possible scenario for the goodness of fit.

The upper panel of Figure 21 presents the cross-sectional mean and interquartile range (IQR) of the R^2 per type, distinguishing between the household sector, individual investors (those with sufficient holdings to estimate separately), and grouped investors[16]. While the model's fit is relatively good for household and individual investors, the fit is less satisfactory for grouped investors. This suggests potential challenges with the current grouping method and highlights non-uniform behavior within these groups. Further breakdowns for the GMM estimation will be presented below. The upper panel indicates that the current approach to grouping investors requires a serious improvement.

The lower panel of Figure 21 reveals an additional severe issue. While fitting the log of the relative portfolio weights works well, fitting the level, which is required for calculating individual portfolio weights $w_{i,t}(n)$ that are e.g. used to compute demand elasticity (2.6) or the market clearing price in variance decomposition, shows a sharp decline in the R^2. This stark loss of predictive power highlights the overall poor performance of the characteristics-based demand equation. Consequently, policy experiments under this esti-

[16]Since there is only one household, there is no IQR.

mation framework lack reliability. Importantly, this analysis is entirely based on in-sample fit, which should generally exhibit stronger performance.

Figures 22, 23 and 24 display the in-sample R^2 for the GMM estimates. A declining trend in all Figures is evident due to the gradual introduction of "zeros" over time. This trend underscores the challenges posed by the "zeros," which significantly diminish the goodness of fit. Particularly for grouped investors, the R^2 values are often negative, implying that using the inverse of the portfolio size, i.e. $1/N$ weights, offers better predictions than the characteristics-based demand equation (2.1). This result poses a significant challenge to demand-based asset pricing, as the computed counterfactual demands used in the variance decomposition are likely to be highly unreliable.

Moreover, for grouped investors positive R^2 are relatively rare, indicating challenges in estimating their behavior meaningfully. To illustrate this, Figure 25 presents the distribution of R^2 across the entire time series. For grouped investors, 60% of cases are better predicted using the average $1/N$ as the portfolio holding, while for investors who can be individually estimated due to holding many assets, over 25% are better fitted with $1/N$. These findings indicate that GMM Estimates do not provide a strong in-sample fit, and the characteristics-based demand equation does not explain portfolio holdings to an extent required to reliably compute counterfactual demands and thus counterfactual returns.

Due to the above, the grouping of investors using their type and AUM is severely insufficient to determine the similarity of their behavior. As a result, the high latent demand components come as no surprise since the GMM Estimation produces negligible explanatory power. Likewise, the inability to estimate grouped investors undermines the findings of KY19 who attribute stock volatility to these grouped investors (see their Table 4).

6.1 Non-Linear Least Squares

We have tried to remedy the insufficient fit of the GMM Estimation by estimating (4.1) using a penalized non-linear least squares (NLLS) estimator with five-fold cross-validation. We use a penalized regression to reduce the potential problem that NLLS could overfit in-sample, as it is a much more flexible estimation framework than GMM. For the penalty term, we use the smooth mean absolute error $tanh(x/2)x$, which for larger x behaves as an L_1-error and is smoothly transitioned to an L_2 error for small x (see Noel et al. (2024) e.g. their Figure 3). A purely L_1-error was not computationally feasible while the computationally much faster L_2 error insufficiently penalized small coefficients. The SMAE, which mitigates the problems of the L_2-error, "only" took several days to estimate and was therefore computationally more feasible than the L_1 error. Hence, for each bin in each cross-section we minimized the following loss

$$Loss_{i,t} = \min_{\boldsymbol{\beta}_{i,t}, \beta_{1,i,t}<1} \sum_{n \in \mathcal{P}_{i,t}} \left(\delta_{i,t}(n) - \widehat{\delta}_{i,t}(n, \boldsymbol{\beta}_{i,t}) \right)^2 + \lambda \cdot tanh\left(\frac{\boldsymbol{\beta}_{i,t}}{2} \right) \boldsymbol{\beta}_{i,t} \quad \forall t \in \mathbb{T},\ \forall i \in \mathbb{B}_t$$

with five-fold cross-validation, where $\widehat{\delta}_{i,t}(\beta)$ is given by (2.1) and $\delta_{i,t}$ is given by (4.6). Furthermore, we once again standardized the explanatory variables and re-transformed them back, just as with the GMM Estimates (see Section A.3). To enforce the restriction on the coefficient of market equity, whenever $\beta_{LNme} \geq 1$, we set it to 0.99 and re-estimated the subset of the other coefficients while keeping $\beta_{LNme} = 0.99$ fixed. This is how e.g. Kuhn-Tucker Optimization can be performed under one inequality constraint.

Figures 26 and 27 display the results for the R^2. As can be seen, the fit improves greatly and in general the characteristics-based demand equation has more explanatory power than the simple $1/N$ prediction. Nonetheless, the overall model performance remains inadequate for policy experiments. Figure 28 displays the regularization parameter, highlighting larger values for grouped investors, confirming their behavioral heterogeneity so that enforcing a uniform $\boldsymbol{\beta}_{i,t}$ is inadequate. Despite the much better performance of the estimation, variance decomposition results mirror those from the baseline GMM Estimation as Table 5 shows. Hence, the characteristics-based demand equation can simply not generate enough fluctuations to explain stock return volatility.

Table 5: Variance Decomposition NLLS Cross Validation

	% of Variance
Supply:	
Shares outstanding	4.0
	(0.69)
Stock characteristics	11.78
	(0.95)
Dividend yield	0.22
	(0.02)
Demand:	
Assets under management	1.43
	(0.14)
Coefficients on characteristics	-3.38
	(0.78)
Latent demand: extensive margin	22.98
	(0.52)
Latent demand: intensive margin	63.38
	(1.16)
Observations	63,520

6.2 Partial Effects

In theory, concerns regarding the model's fit of relative portfolio weights might seem unwarranted if one is solely interested in partial effects. For example, in the literature most emphasis has been put on the demand elasticity (2.6). While the validity of demand elasticity estimates can rightly be questioned based on the preceding findings, we further demonstrate that the GMM Estimation yields biased partial effects which we can show

by simulating the data-generating process. The procedure is explained in the remainder of this section.

Firstly, we take the $\{\boldsymbol{\beta}_{i,t}\}_{\forall\, t\in\mathbb{T}, i\in\mathbb{I}_t}$ obtained from the GMM estimation as the true parameter values. Naturally, we could pick any arbitrary values, but it makes sense to use the previously estimated $\boldsymbol{\beta}$ which at least reflect a marginal underlying structure as opposed to picking completely arbitrary values.

Next, we use (4.6) along with the observed relative portfolio weights, $\delta_{i,t}$, to compute $\varepsilon_{i,t}(n)$ such that

$$\varepsilon_{i,t}(n) = \begin{cases} \delta_{i,t}(n) exp\left\{-\boldsymbol{x}_t'(n)\boldsymbol{\beta}_{i,t} - \widetilde{x}_t(n)\right\}, & \forall n \in \mathcal{P}_{i,t} \\ 0, & \forall n \in \mathcal{N}_{i,t} \setminus \mathcal{P}_{i,t} \end{cases} \tag{6.1}$$

where once again due to the grouping we have that $\boldsymbol{\beta}_{i,t} = \boldsymbol{\beta}_{j,t}\ \forall i, j \in B_{g,t}$.

Next, we compute the variance of the *strictly positive* computed error components in (6.1) in order to simulate new error components with appropriate moments. We use the strictly positive error components since these are robust with respect to the construction of the "zeros". We denote this by variance $\sigma^2_{i,t|data} = V(\varepsilon_{i,t}|\varepsilon_{i,t} > 0)$.

Now, we are ready to simulate new error components. For the simulation we pick a log-normal distribution and set $\varepsilon_{i,t}(n)$ to be completely exogenous so that no IV estimation is required. We draw from the log-normal distribution such that $\mathbb{E}[\varepsilon_{i,t}] = 1$ and $V(\varepsilon_{i,t}) = \sigma^2_{i,t|data}$. This is possible as the mean and variance of a log-normal distribution are analytical functions of the two input parameters so that two equations can match two moments exactly. Thus, for each bin $g \in B_{g,t}$ we simulate new error components according to

$$\varepsilon^{sim}_{i,t}(n) \overset{i.i.d}{\sim} \text{Log-Normal}(\mu_{i,t}, \sigma_{i,t}) \qquad s.t. \quad \mathbb{E}[\varepsilon_{i,t}] = 1 \text{ and } V(\varepsilon_{i,t}) = \sigma^2_{i,t|data}.$$

Next, we enforce the "zeros". To do so, we use the constructed share of "zeros" in the bin to randomly set components of $\varepsilon_{i,t}$ to zero with an i.i.d biased coin-flip. For instance, if a bin consists of 1,000 observations, we simulate the strictly positive vector $\varepsilon^{sim}_{i,t}$ of dimension 1000×1 according to the above log-normal distribution. If 30% of the observations are "zeros", for each element of the vector $\varepsilon^{sim}_{i,t}$ we use an i.i.d coin flip that has a 30% chance to change that element to zero and else-wise leave it untouched. Since everything is i.i.d, the distribution of the remaining strictly positive $\varepsilon^{sim}_{i,t}(n)$ is left in tact, in particular they still have mean 1.

Finally, we simulate new relative portfolio weights, $\delta^{sim}_{i,t}(n)$, by simply using the characteristics-based demand equation

$$\delta^{sim}_{i,t}(n) = exp\left\{\boldsymbol{x}_t'(n)\boldsymbol{\beta}_{i,t} + \widetilde{x}_t(n)\right\} \varepsilon^{sim}_{i,t}(n).$$

We can now put GMM to the test. We use $\{\delta^{sim}_{i,t}, \boldsymbol{x}_t\}$ to estimate $\boldsymbol{\beta}_{i,t}$ as before using (4.6).

We omit the IV since we know the true data-generating process and we simulated the error component to be completely exogenous. Furthermore, we do not impose the restriction since the input $\boldsymbol{\beta}$ to generate the data generating process fulfilled this requirement already. On top of that, there is no adverse affect from grouping as the investors within the group (bin) exhibit homogeneous behavior. In short, the conditions are configured to be as favorable as possible for GMM to deliver accurate results.

Using the estimated $\widehat{\boldsymbol{\beta}}_{i,t}$ from the simulated data, we compute the demand elasticity via (2.6) and compare it to the true demand elasticity, as shown in Figure 29. The results unequivocally show that the demand elasticities derived from GMM are biased. This is a critical failure, especially given that the bias persists even under the most ideal conditions. It is worth noting that in the initial period, where the proportion of "zeros" is 0%, GMM aligns with OLS and correctly estimates the demand elasticity. However, this scenario is unrealistic, as in practice, such estimates would likely suffer from selection bias (compare KY19 Figure F1). Given these results, the magnitude of bias in real world applications is expected to be significantly worse, rendering the demand elasticity estimates produced by GMM unreliable. These findings raise significant concerns about the reliability of such estimates, indicating that caution should be exercised when interpreting them for policy experiments.

6.3 Constant Investment Universe

An evident source for latent demand is the changing investment universe. It is not modeled by e.g. a probit or logit whether an investor will select a stock or not in the next period. Therefore, by design, the change in the investment universe is exogenous and a source of latent demand. To assess the effect, we kept the investment universe constant from one period to the next for the variance decomposition. Hence, for each time tuple t and $t+4$ we consider two alterations. Firstly, we only consider managers who exist in both periods. Secondly, we restrict the investment universe to the intersection of stocks of the overlapping managers, i.e. the augmented investment universe of each overlapping manager is given by $\widetilde{\mathcal{P}}_{i,t} = \widetilde{\mathcal{P}}_{i,t+4} = \mathcal{P}_{i,t} \cap \mathcal{P}_{i,t+4}$. Thus, each manager exists in both periods and holds the same set of stocks in both periods.

As this cuts out stocks, for each $n \in \bigcup_i \widetilde{\mathcal{P}}_{i,t}$ a new initial market clearing price must be computed. Otherwise, the variance decomposition works exactly as described in Section 2.3. Table 6 displays the results of the variance decomposition. As can be seen, the model breaks down and produces nonsensical results. This can be seen as an indictment to DBAP, but at the very least, it is clear that the modeling of the investment universe requires more study to fully assess its impact on the framework.

Table 6: Variance Decomposition Constant Investment Universe

	% of Variance
Supply:	
Shares outstanding	1.07
	(1.37)
Stock characteristics	-1.14
	(3.0)
Dividend yield	0.0
	(0.0)
Demand:	
Assets under management	-1.05
	(2.18)
Coefficients on characteristics	61.56
	(46.81)
Latent demand	39.56
	(50.6)
Observations	63,520

7 Conclusion

This paper introduced a rigorous notation and explained KY19's framework in full detail, along with providing an extensively commented and flexible Python library, thereby removing barriers to entry. Our analysis showed that KY19's framework cannot explain the variance of stock returns. There are several reasons for this. First, many investors hold concentrated portfolios, so they must be grouped together, but the selection criteria chosen by KY19 results in non-homogeneous behavior of the group members. This classification problem is not easy to solve, as it is not only high-dimensional, but it is also not easy to obtain data to further classify investors into finer categories. Second, variable selection has shown that in the KY19 framework, investors care only about a stock's price and little else. Third, attempting to address the selection bias by introducing "zeros" leads to serious estimation problems, including biased partial effects. Further research could focus on refining investor preferences. For example, the $\boldsymbol{\beta}$ coefficients might also depend on the type of stock, i.e. investors might value certain characteristics more for technology companies than for pharmaceutical companies. In addition, characteristics other than accounting measures, such as sentiment, may play a large role in explaining investor demand. Future research could also implement a stock selection model to predict whether and why an investor would or would not select a particular stock to endogenize the investment universe. Recent advances in large language models could potentially help to classify investors with much more precision by analyzing company reports, since e.g. the classification of "investment advisor" by KY19 is very broad. This could mitigate the estimation problems caused by grouping.

References

Anderson, A. & Brockman, P. (2018), 'An examination of 13f filings', *Journal of Financial Research* **41**(3), 295–324.

Backus, M., Conlon, C. & Sinkinson, M. (2021), 'Common ownership in america: 1980–2017', *American Economic Journal: Microeconomics* **13**(3), 273–308.

Bahaj, S., Czech, R., Ding, S. & Reis, R. (2023), 'The Market for Inflation Risk', *SSRN Working Paper 4488881* .

Ben-David, I., Franzoni, F., Moussawi, R. & Sedunov, J. (2021), 'The granular nature of large institutional investors', *Management Science* **67**(11), 6629–6659.

Caner, M. (2009), 'Lasso-type gmm estimator', *Econometric Theory* **25**(1), 270–290.

Caner, M. & Zhang, H. H. (2014), 'Adaptive elastic net for generalized methods of moments', *Journal of Business & Economic Statistics* **32**(1), 30–47.
URL: *https://doi.org/10.1080/07350015.2013.836104*

Chen, A. Y. & Zimmermann, T. (2022), 'Open source cross-sectional asset pricing', *Critical Finance Review* **27**(2), 207–264.

Cochrane, J. H. (2011), 'Presidential Address: Discount Rates', *Journal of Finance* **66**(4), 1047–1108.

Fama, E. F. (1970), 'Efficient capital markets: A review of theory and empirical work', *The Journal of Finance* **25**(2), 383–417.

Fuchs, W. M., Fukuda, S. & Neuhann, D. (2023), 'Demand-system asset pricing: Theoretical foundations', *SSRN Working Paper 4672473* .

Green, J., Hand, J. R. M. & Zhang, X. F. (2017), 'The characteristics that provide independent information about average u.s. monthly stock returns', *The Review of Financial Studies* **30**(12), 4389–4436.
URL: *https://doi.org/10.1093/rfs/hhx019*

Greene, W. (2019), *Econometric Analysis Global Edition*, Pearson.

Gu, S., Kelly, B. & Xiu, D. (2020), 'Empirical asset pricing via machine learning', *The Review of Financial Studies* **33**(5), 2223–2273.
URL: *https://doi.org/10.1093/rfs/hhaa009*

Haddad, V., Huebner, P. & Loualiche, E. (2025), 'How Competitive is the Stock Market? Theory, Evidence from Portfolios, and Implications for the Rise of Passive Investing', *American Economic Review (forthcoming)* .

Han, X., Roussanov, N. L. & Ruan, H. (2021), 'Mutual fund risk shifting and risk anomalies', *SSRN Working Paper 3931449* .

Hansen, L. P. (1982), 'Large Sample Properties of Generalized Method of Moments Estimators', *Econometrica* **50**(4), 1029.

Koijen, R. S. J., Richmond, R. J. & Yogo, M. (2023), 'Which investors matter for equity valuations and expected returns?', *Review of Economic Studies* **91**(4), 2387–2424.

Koijen, R. S. J. & Yogo, M. (2019), 'A demand system approach to asset pricing', *Journal of Political Economy* **127**(4), 1475–1515.

Kozak, S., Nagel, S. & Santosh, S. (2020), 'Shrinking the cross-section', *Journal of Financial Economics* **135**(2), 271–292.
URL: *https://www.sciencedirect.com/science/article/pii/S0304405X19301655*

Leng, C., Lin, Y. & Wahba, G. (2006), 'A note on the lasso and related procedures in model selection', *Statistica Sinica* **16**(4), 1273–1284.
URL: *http://www.jstor.org/stable/24307787*

Li, J. (2022), 'What drives the size and value factors?', *The Review of Asset Pricing Studies* **12**(4), 845–885.

Newey, W. & McFadden, D. (1994), 'Large sample estimation and hypothesis testing', *Handbook of Econometrics* **4**, 2111–2245.

Noel, M., Banerjee, A., Oswal, Y., Amali, G. & Muthiah-Nakarajan, V. (2024), 'Alternate loss functions for classification and robust regression can improve the accuracy of artificial neural networks'.
URL: *https://arxiv.org/abs/2303.09935*

Sammon, M. & Shim, J. (2024), 'Who Clears the Market When Passive Investors Trade?', *SSRN Working Paper 4777585* .

Shiller, R. J. (1981), 'Do stock prices move too much to be justified by subsequent changes in dividends?', *The American Economic Review* **71**(3), 421–436.

van der Beck, P. (2022*a*), 'Flow-Driven ESG Returns', *Swiss Finance Institute Research Paper Series, No. 21-71* .

van der Beck, P. (2022*b*), 'On the Estimation of Demand-Based Asset Pricing Models', *Swiss Finance Institute Research Paper Series, No. 21-71* .

Wharton Research Data Services (2017), 'Research Note Regarding Thomson-Reuters Ownership Data Issues'.

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

Figure 1: Main Mechanism DBAP

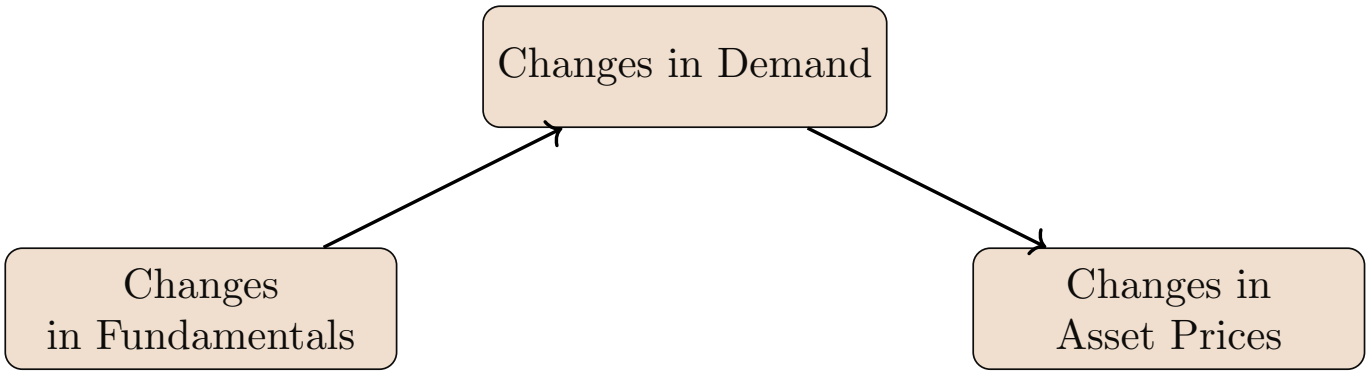
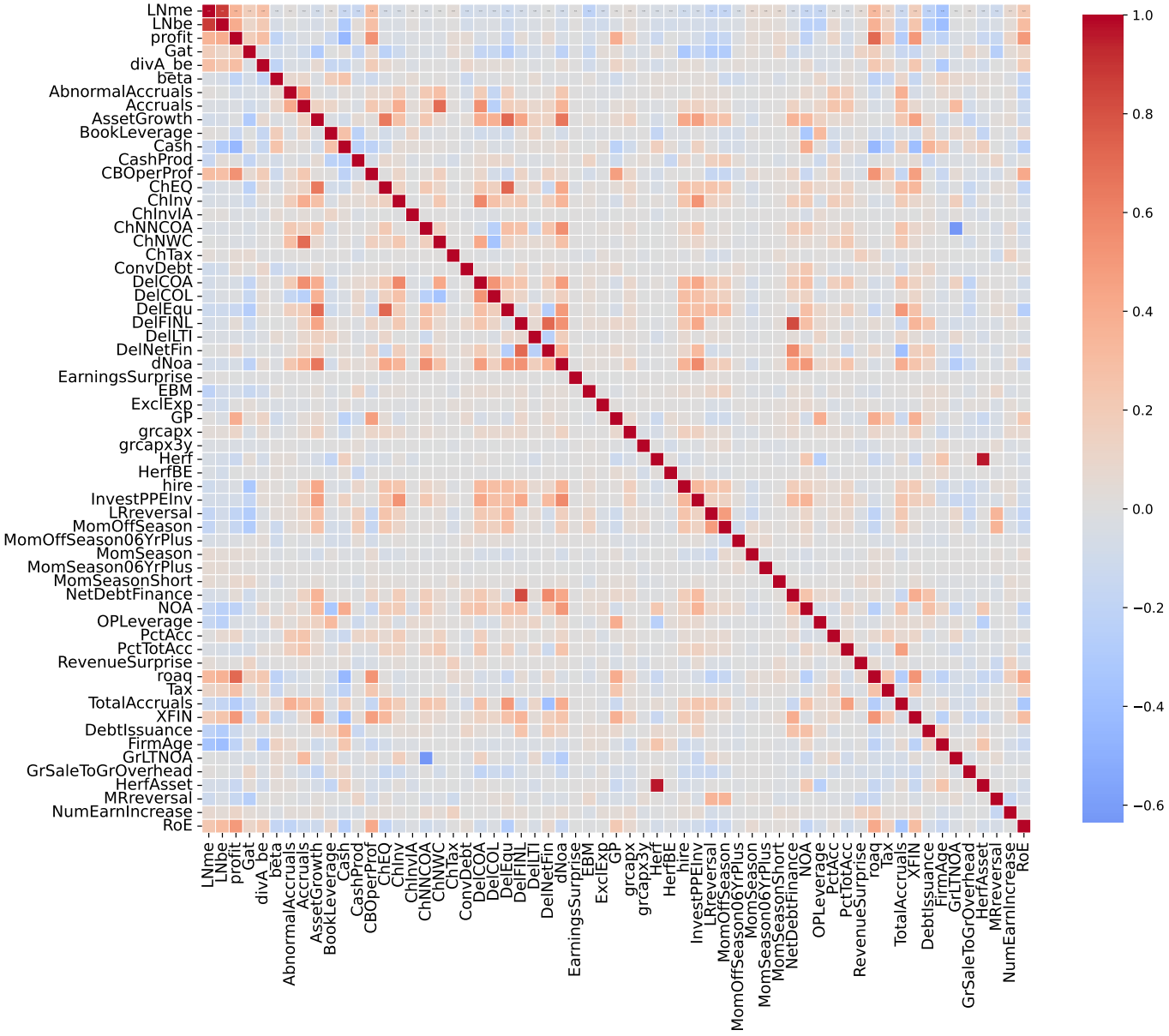
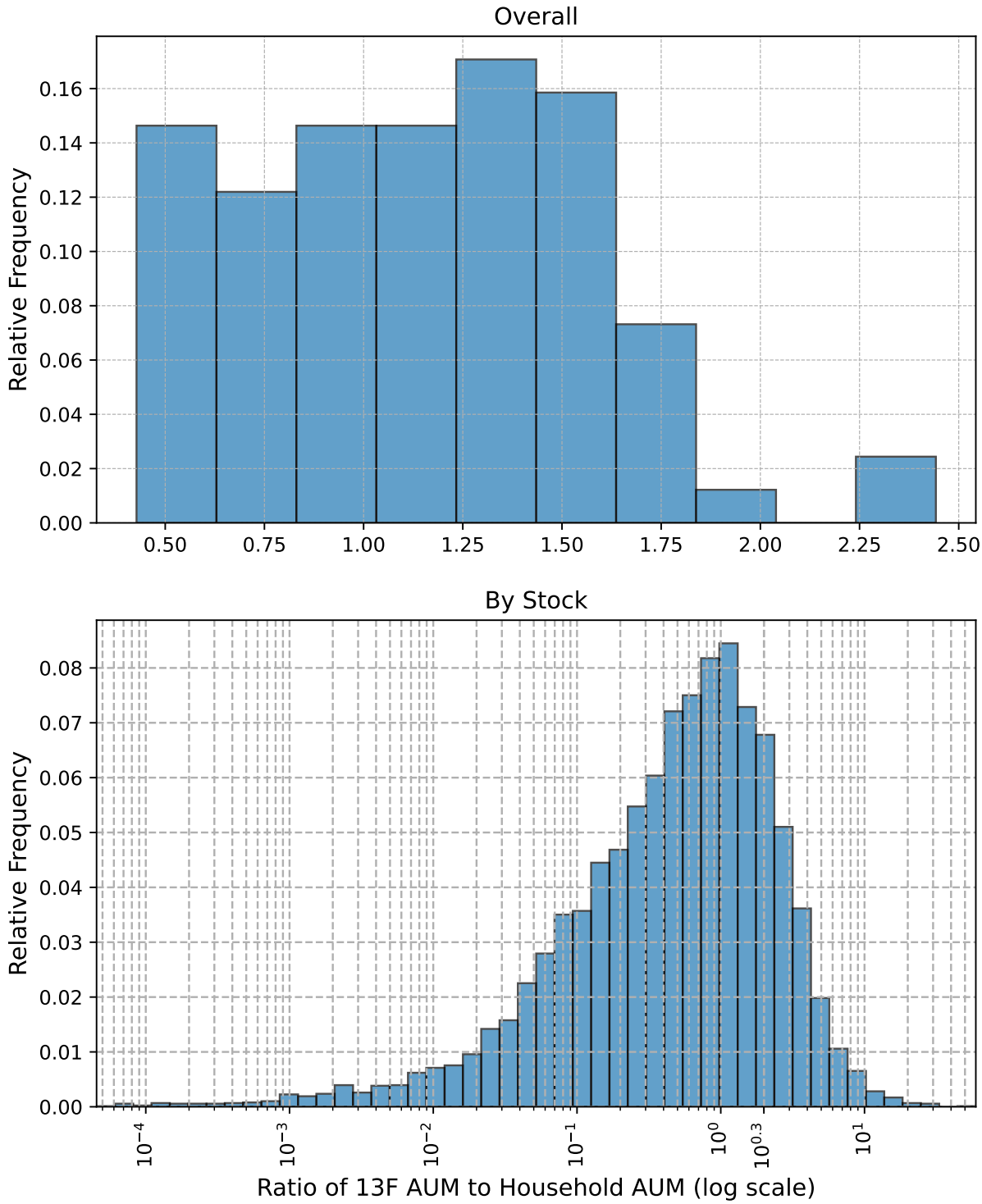


Figure 2: Correlation Matrix of Features



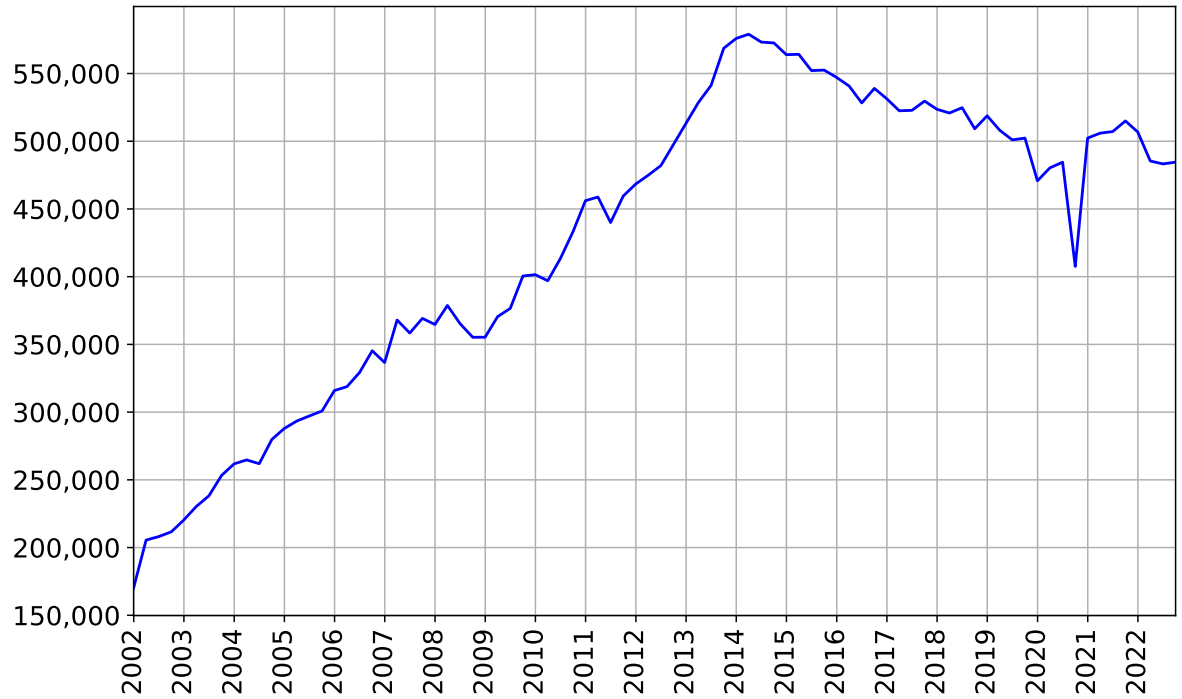
Displays the correlation matrix over the entire sample period for all the features. Baseline characteristics of Koijen & Yogo (2019) are in the top left corner. Acronyms are as in Chen & Zimmermann (2022).

Figure 3: Ratio of 13F to Household AUM



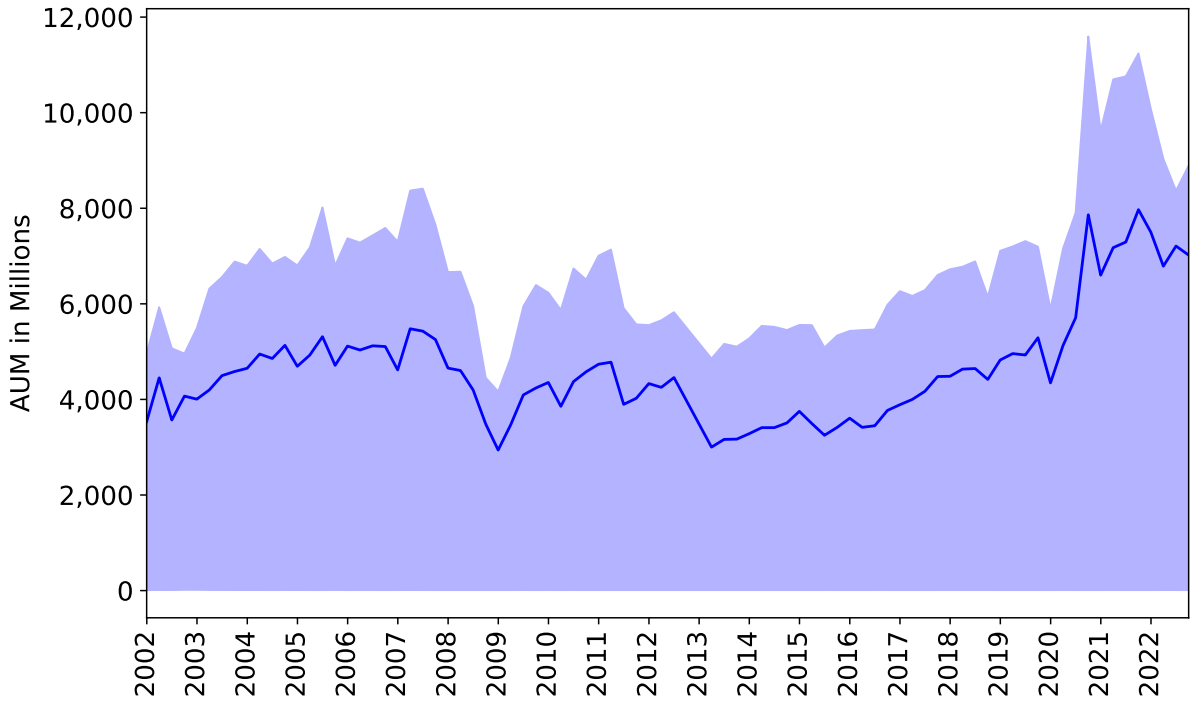
Displays the histogram of the ratio of aggregated 13F AUM to Household AUM. Upper panel contains the overall quarterly ratio. For the lower panel the ratio was computed per stock in the cross-section and then averaged over the entire time series.

Figure 4: Number of Observations



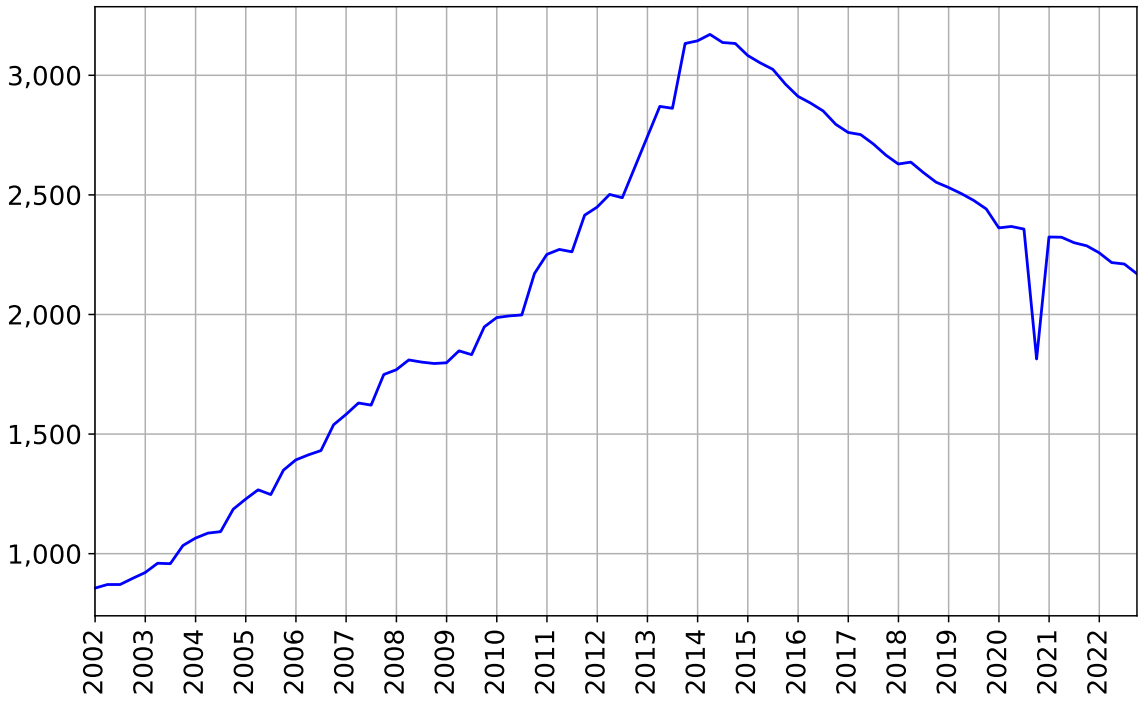
Displays the number of observations in the dataset per quarter.

Figure 5: AUM 13F Investors.



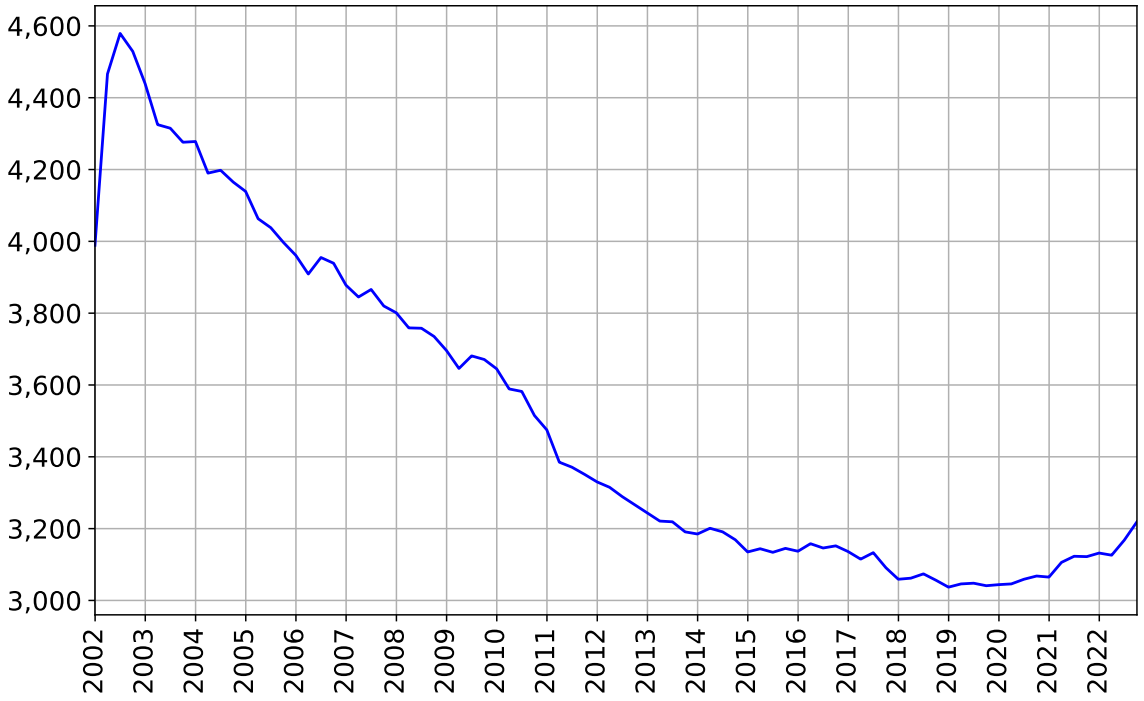
Displays the quarterly AUM of the 13F investors. Solid line is the cross-sectional median. Shaded Area is the cross-sectional bottom 90%.

Figure 6: Number of Managers



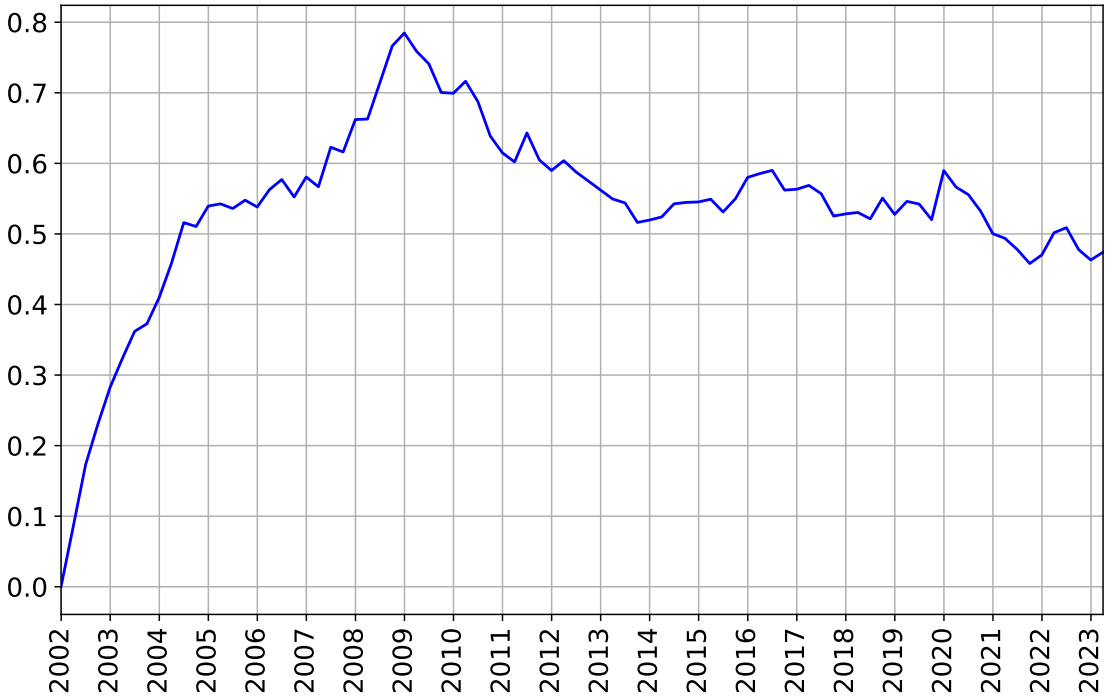
Displays the number of managers per quarter.

Figure 7: Number of Stocks



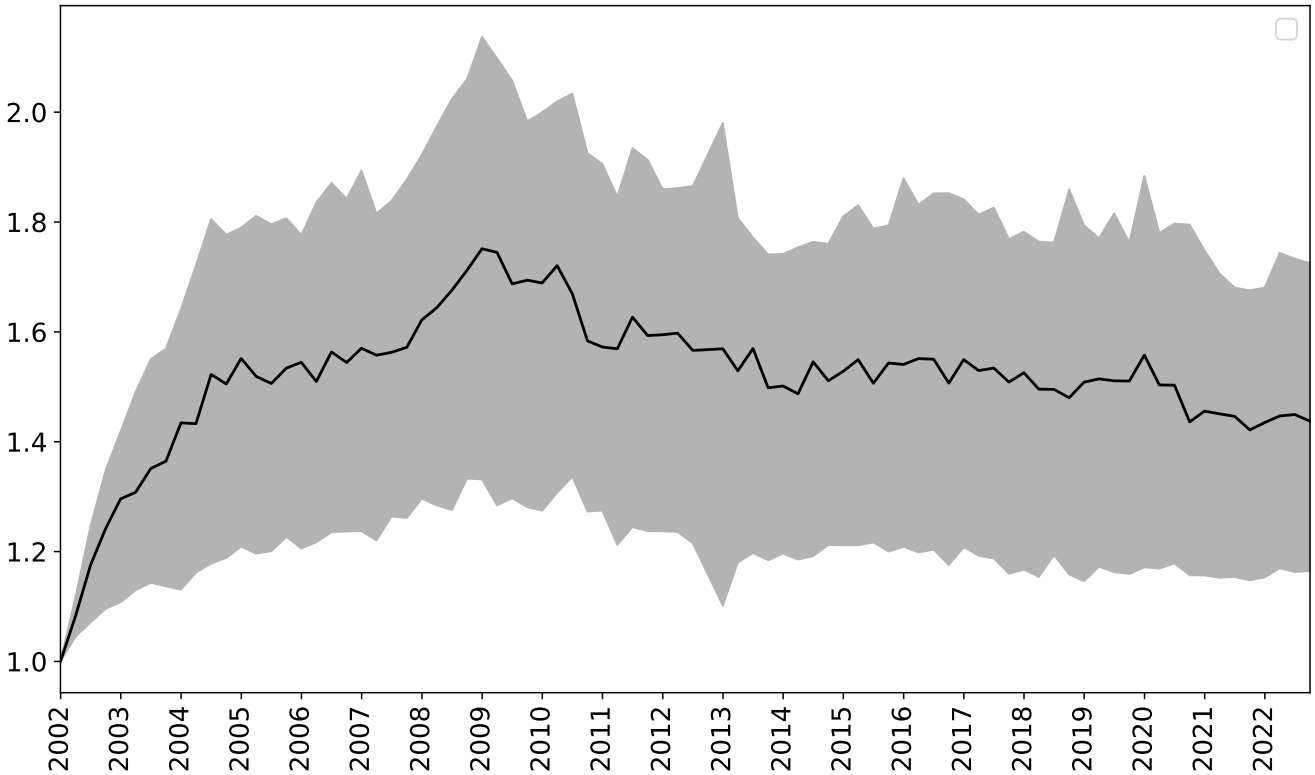
Displays the number of stocks held at least by one 13F investor per quarter.

Figure 8: Share of "Zeros"



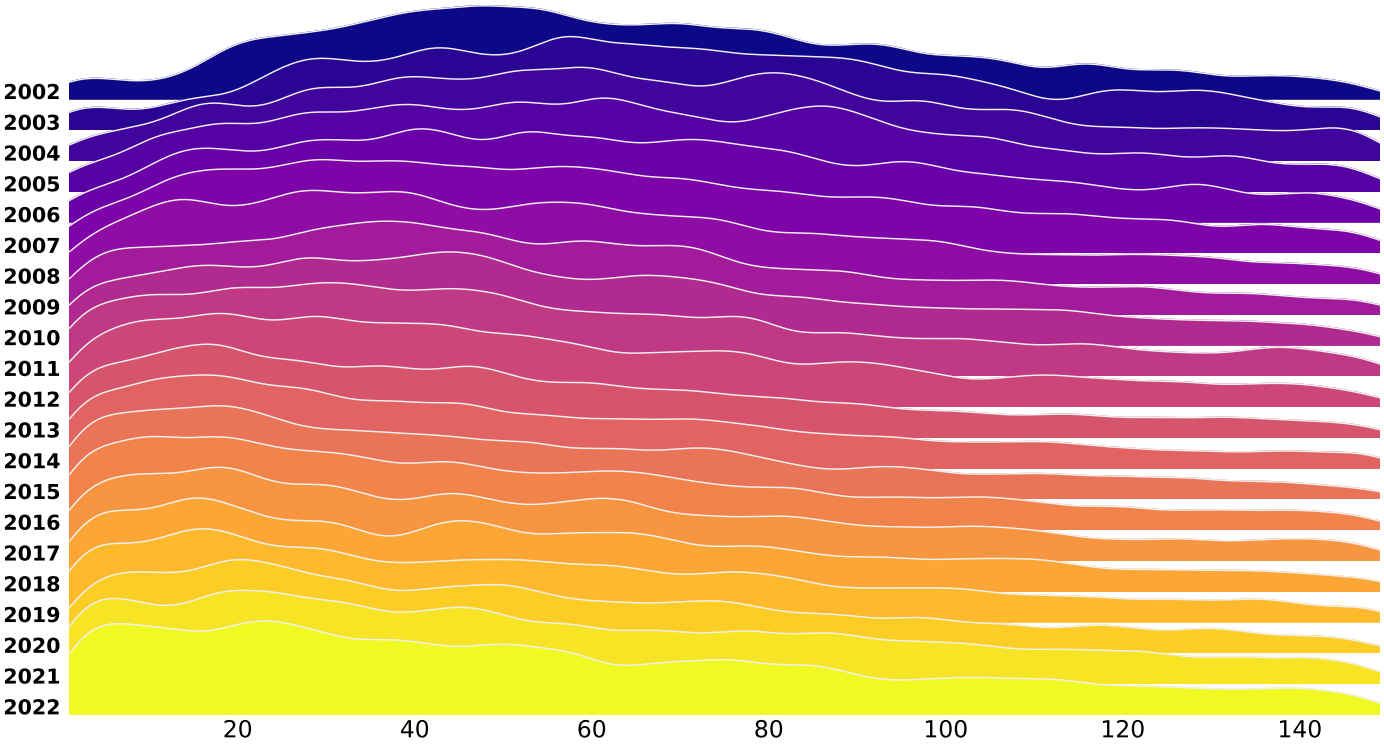
Displays the share of "zeros" of overall portfolio holdings per quarter.

Figure 9: Epsilon Bias



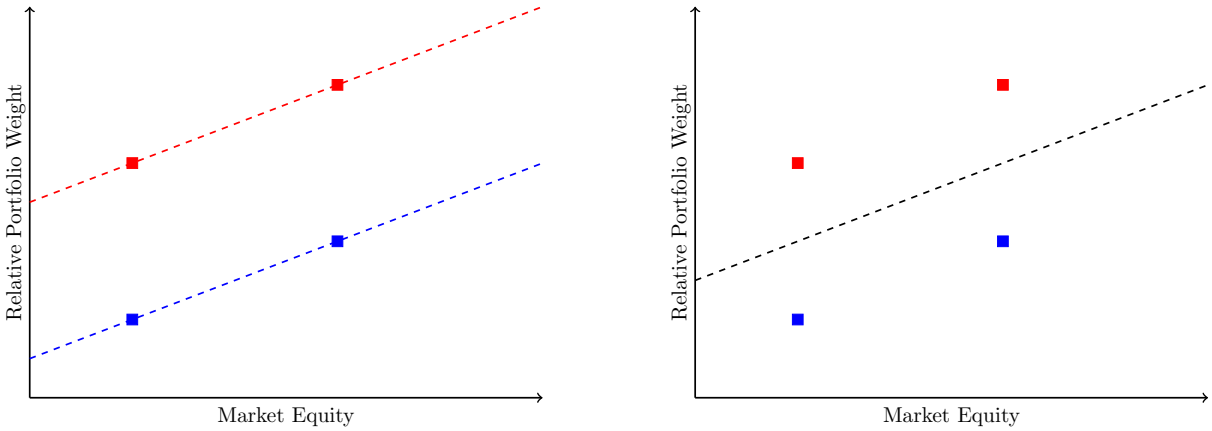
Displays the median value of the averages of strictly positive epsilon values within each bin. In other words, first calculate the average of the positive epsilon values for each bin separately, and then take the median of these averages across all bins. Shaded Area is the interquartile range of averages.

Figure 10: Number of Holdings per Investor



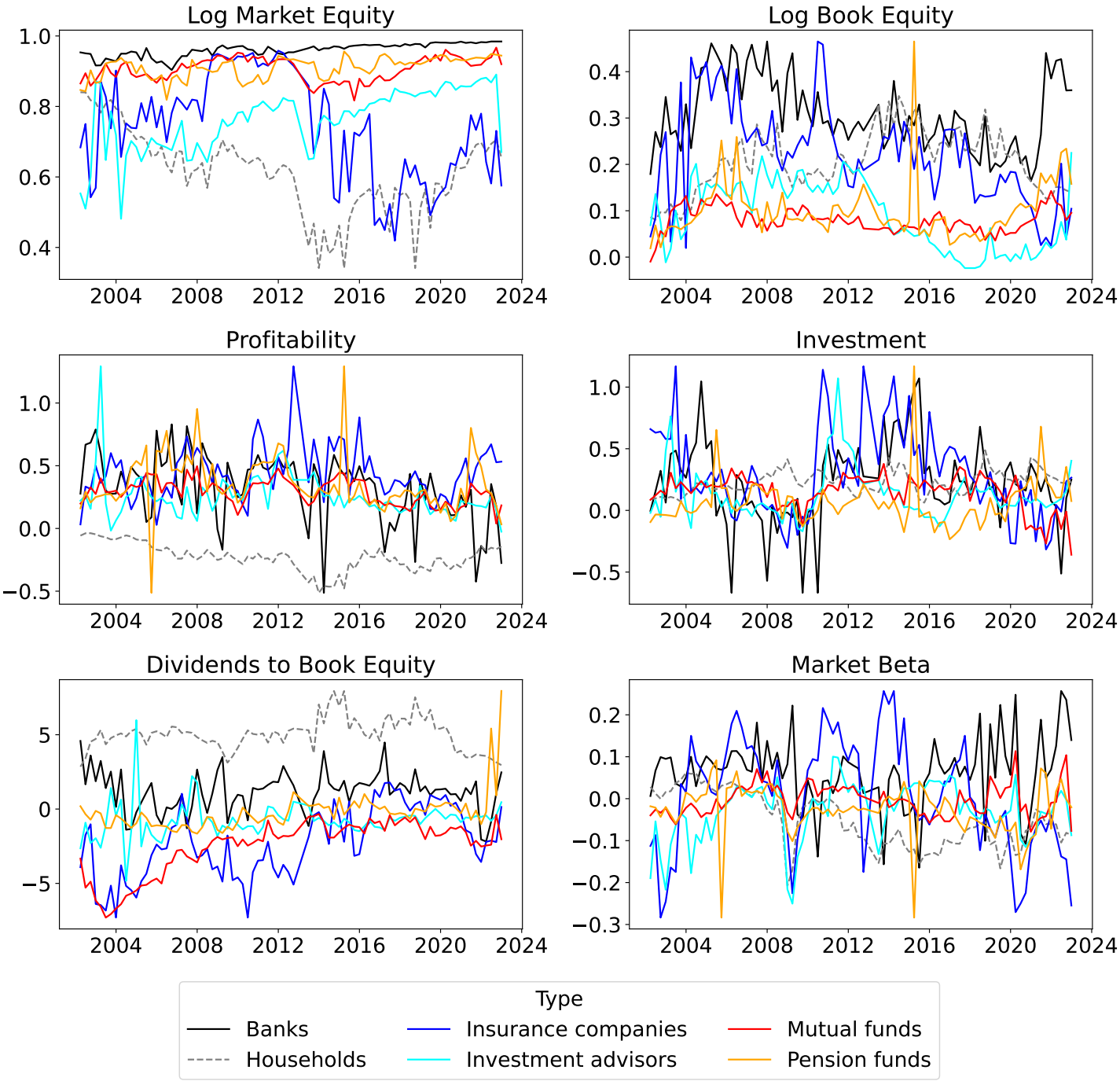
Truncated Histogram of number of strictly positive holdings per investor. Values at the right tail diminish uniformly and are therefore cut off for enhanced visibility.

Figure 11: Fixed Effect



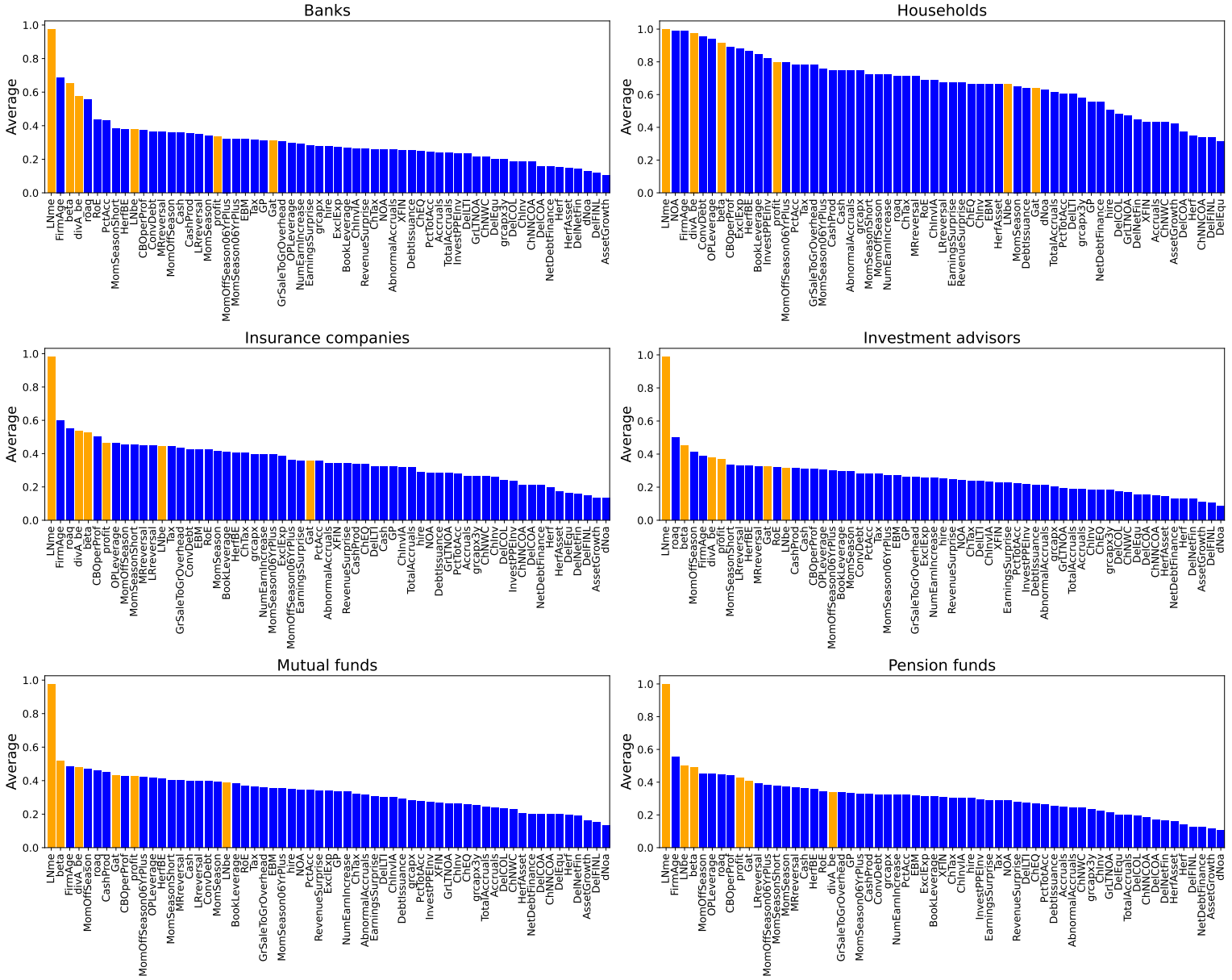
Visualises the benefit of the fixed effect. Squares are portfolio holdings to fit.

Figure 12: Time Series of Coefficients



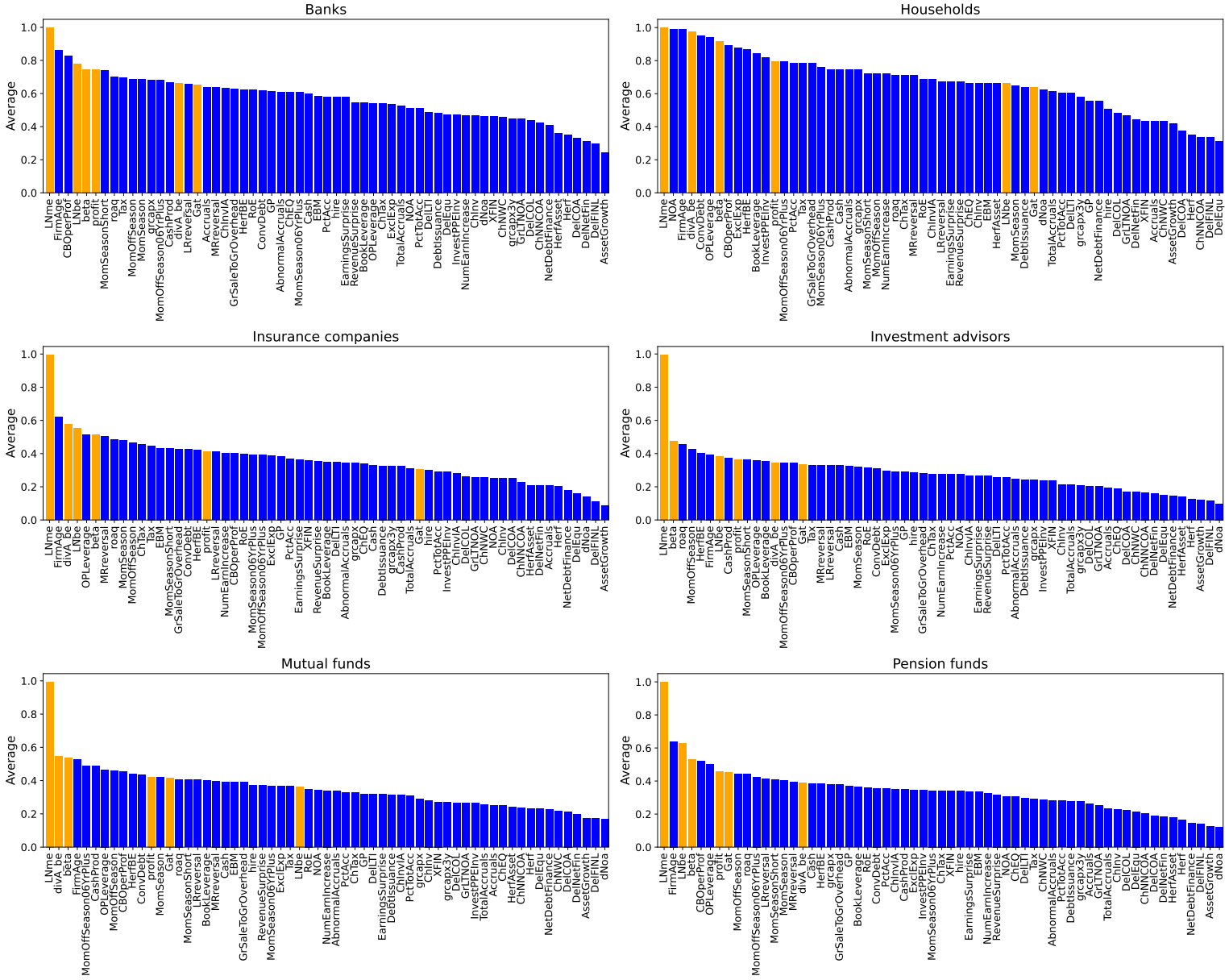
This Figure reports the cross-sectional mean of the estimated coefficients by institution type, weighted by assets under management. Compare with Figure 3 in Koijen & Yogo (2019).

Figure 13: LASSO Variable Selection Pooled



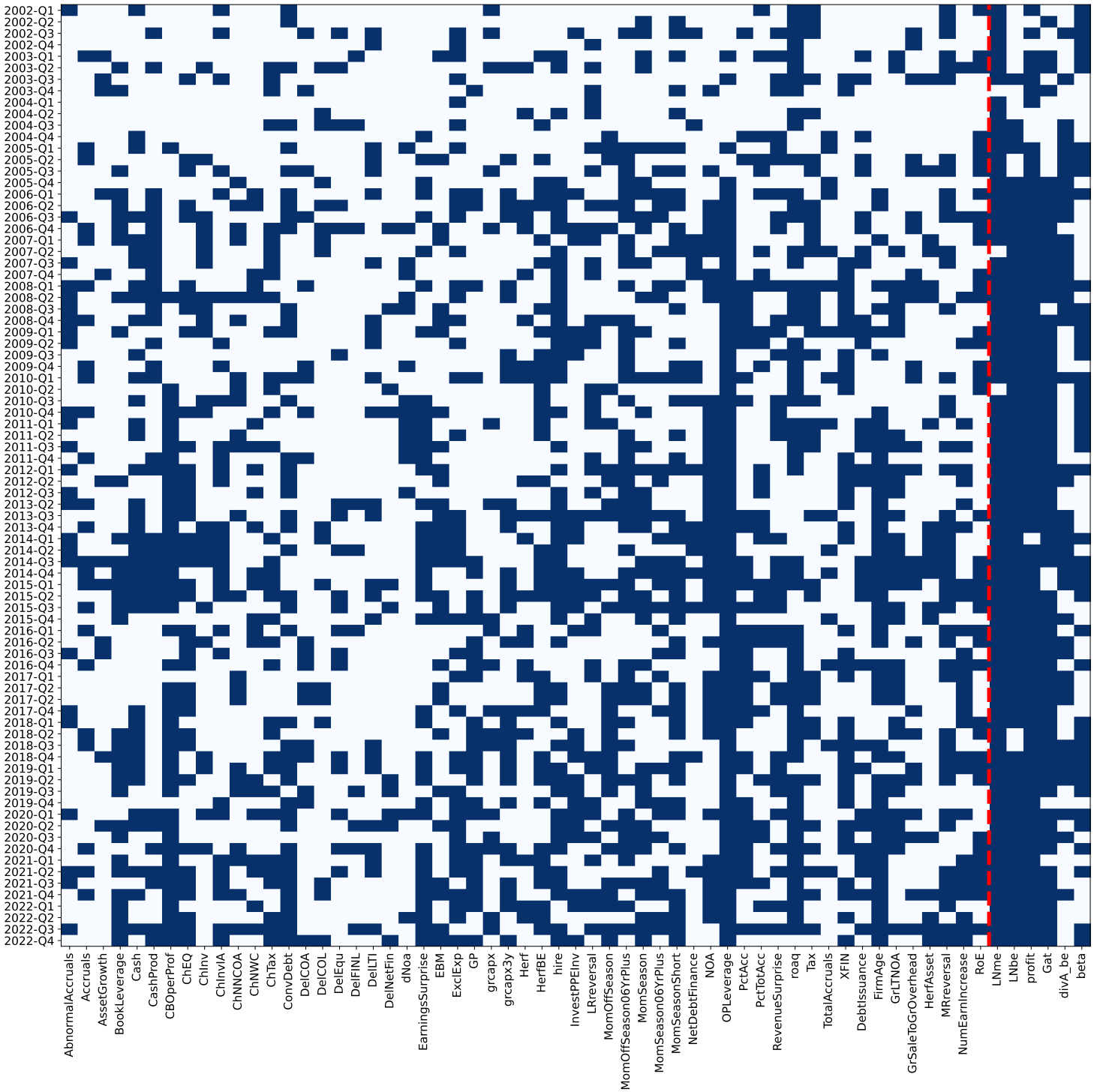
Displays how often a characteristic was chosen per type. Acronyms are as in Chen & Zimmermann (2022). Calculation is pooled over all investors and all quarters per type. Characteristics from the original Koijen & Yogo (2019) are displayed in orange.

Figure 14: LASSO Variable Selection Valued Weighted



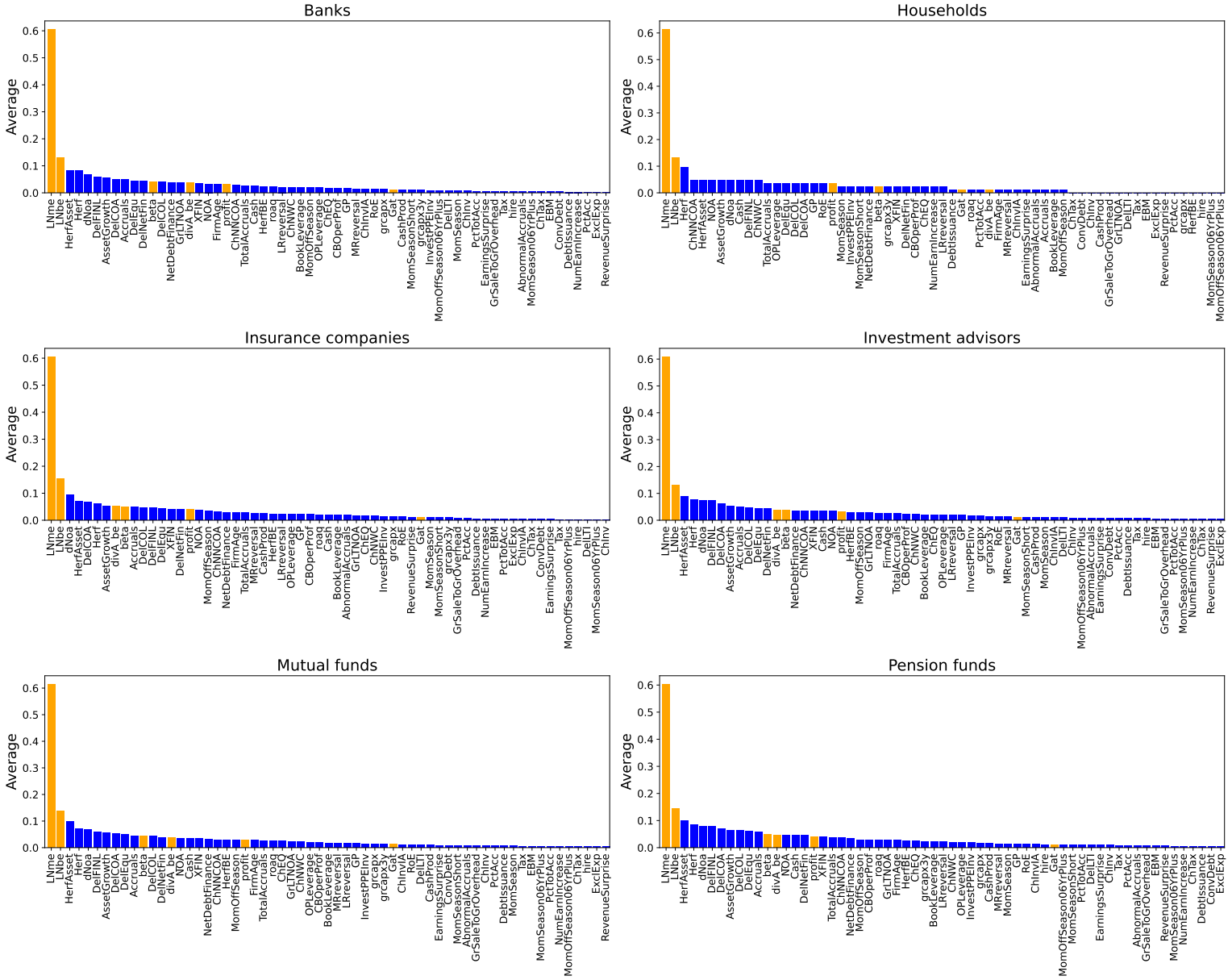
Displays how often a characteristic was chosen per type. Acronyms are as in Chen & Zimmermann (2022). Averages are valued weighted within years and within types. Per type, over all years an unweighted average is used. Characteristics from the original Koijen & Yogo (2019) are displayed in orange.

Figure 15: LASSO Variable Selection Results AQR Capital Management



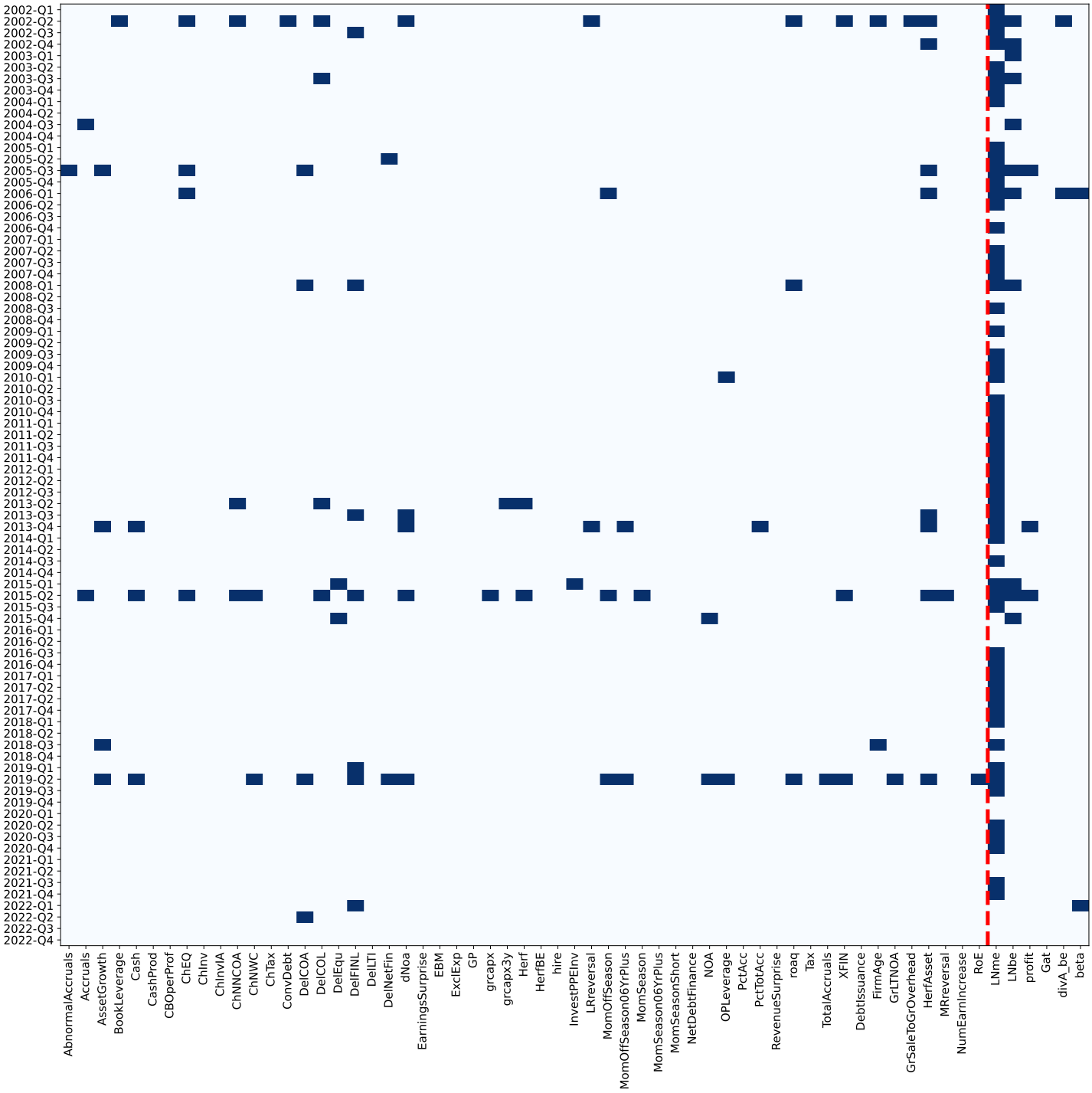
Displays how often the LASSO selected a characteristic (dark blue pixels) for the investor at each time point. Light blue pixels signal that the characteristic was not chosen. Koijen & Yogo (2019) baseline characteristics are to the right of the dotted red line. Acronyms are as in Chen & Zimmermann (2022).

Figure 16: Adaptive LASSO Variable Selection Pooled - OLS Weights



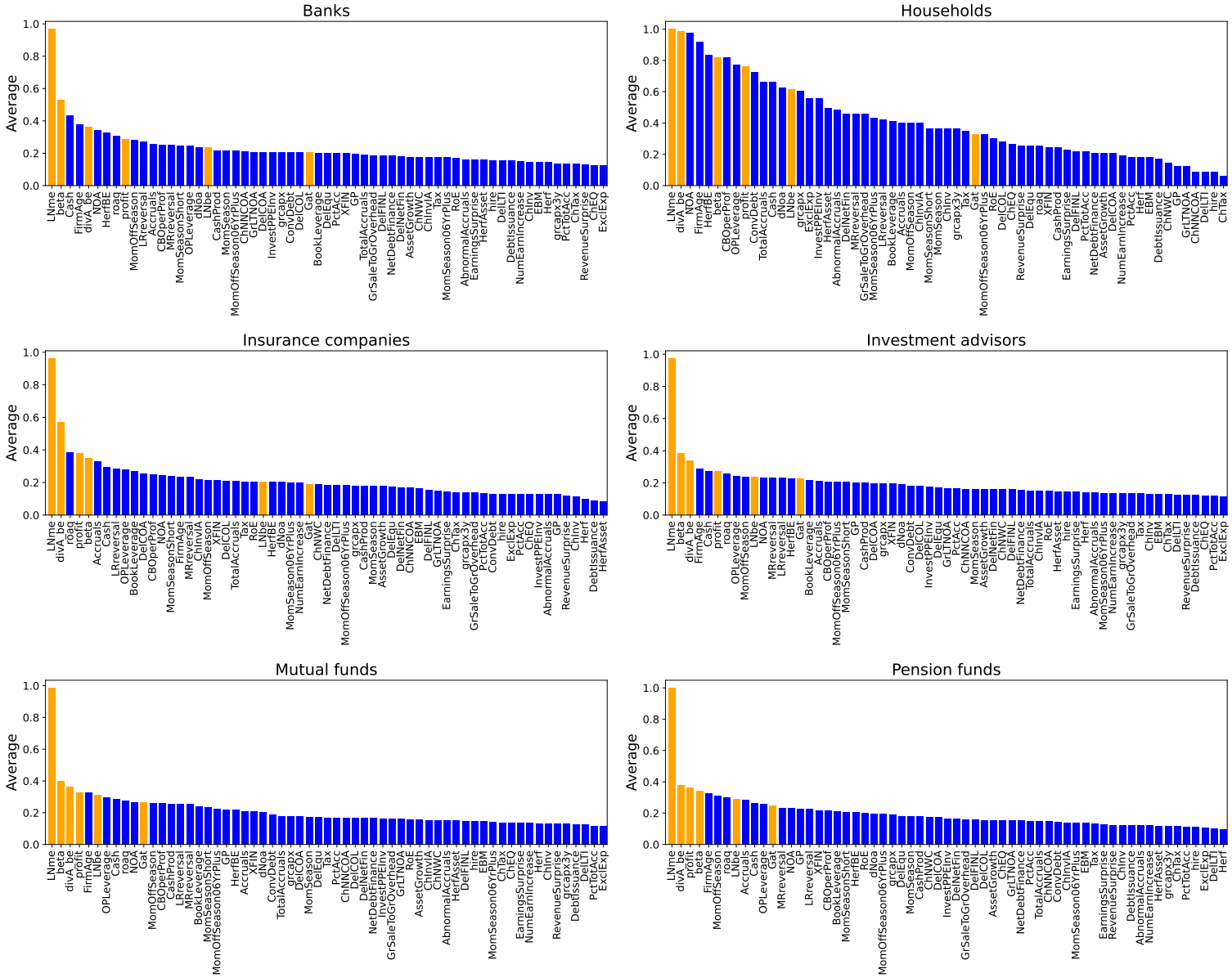
Displays how often a characteristic was chosen per type. Acronyms are as in Chen & Zimmermann (2022). Calculation is pooled over all investors and all quarters per type. Characteristics from the original Koijen & Yogo (2019) are displayed in orange.

Figure 17: Adaptive LASSO Variable Selection Results AQR Capital Management



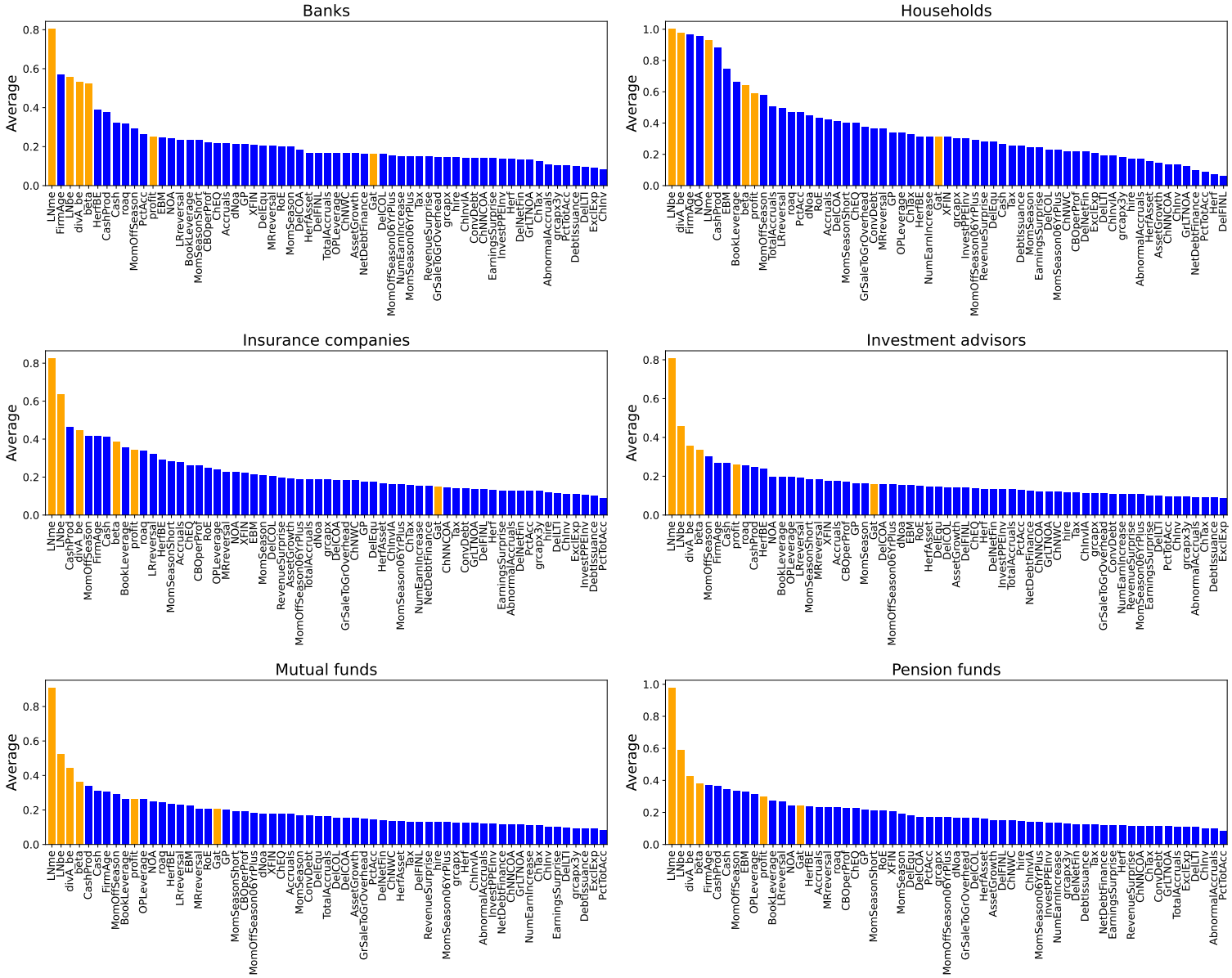
Displays how often the adaptive LASSO selected a characteristic (dark blue pixels) for the investor at each time point. Light blue pixels signal that the characteristic was not chosen. Koijen & Yogo (2019) baseline characteristics are to the right of the dotted red line. Acronyms are as in Chen & Zimmermann (2022).

Figure 18: Backward Selection GMM



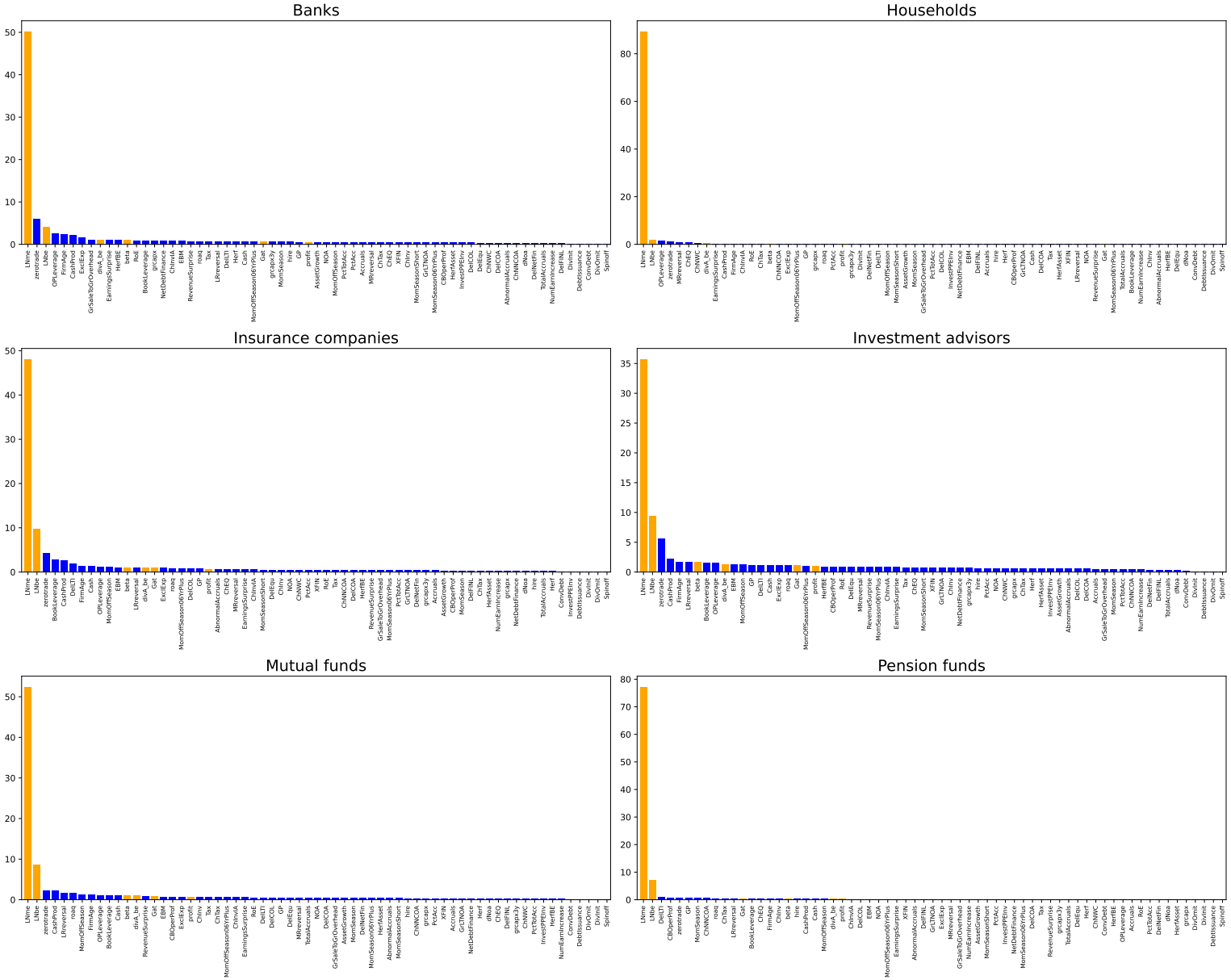
Displays how often a characteristic was chosen per type. Acronyms are as in Chen & Zimmermann (2022). Calculation is pooled over all investors and all quarters per type. Characteristics from the original Koijen & Yogo (2019) are displayed in orange.

Figure 19: Backward Selection IV2SLS



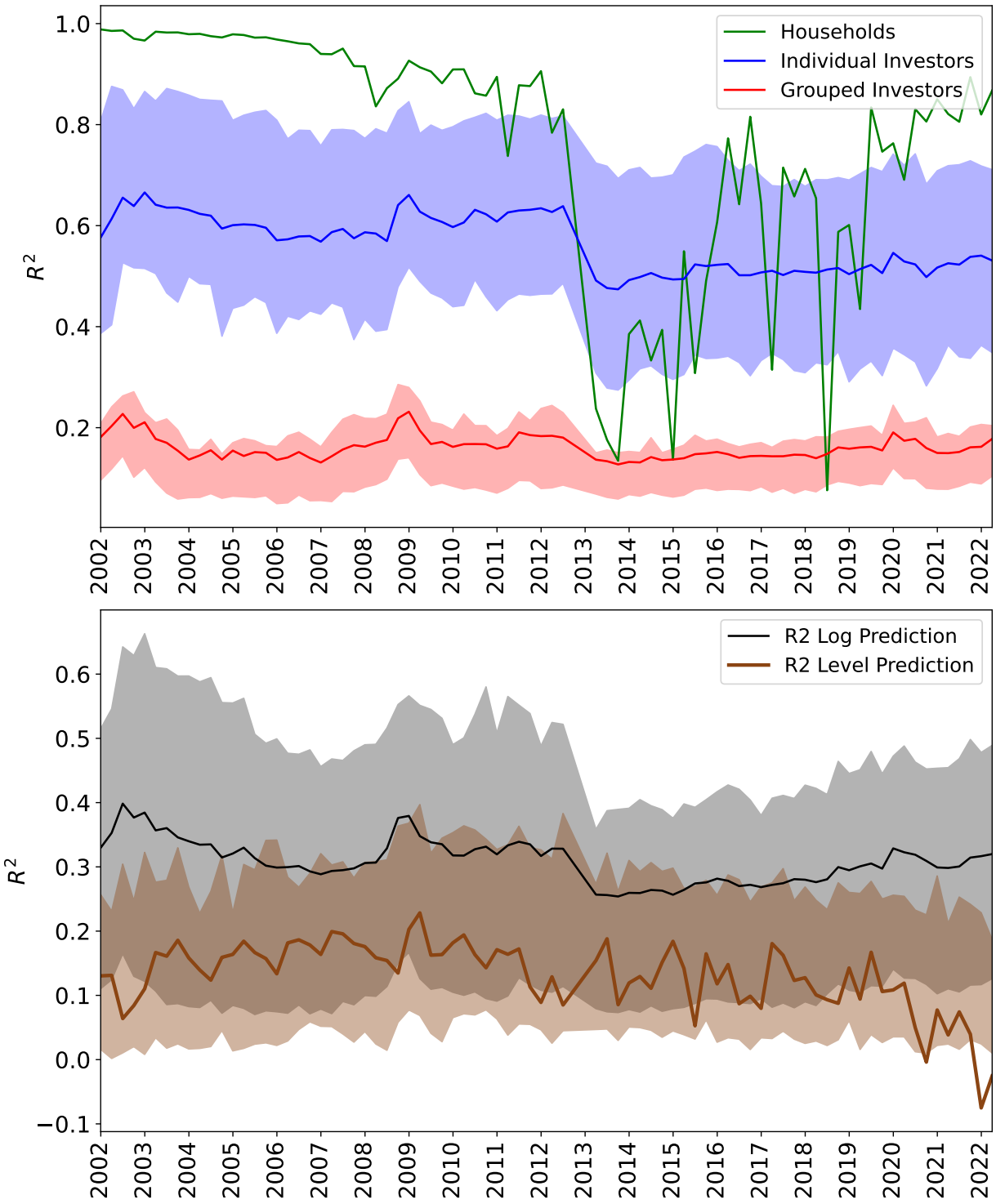
Displays how often a characteristic was chosen per type. Acronyms are as in Chen & Zimmermann (2022). Calculation is pooled over all investors and all quarters per type. Characteristics from the original Koijen & Yogo (2019) are displayed in orange.

Figure 20: Variable Importance in Gradient Boosting Regression



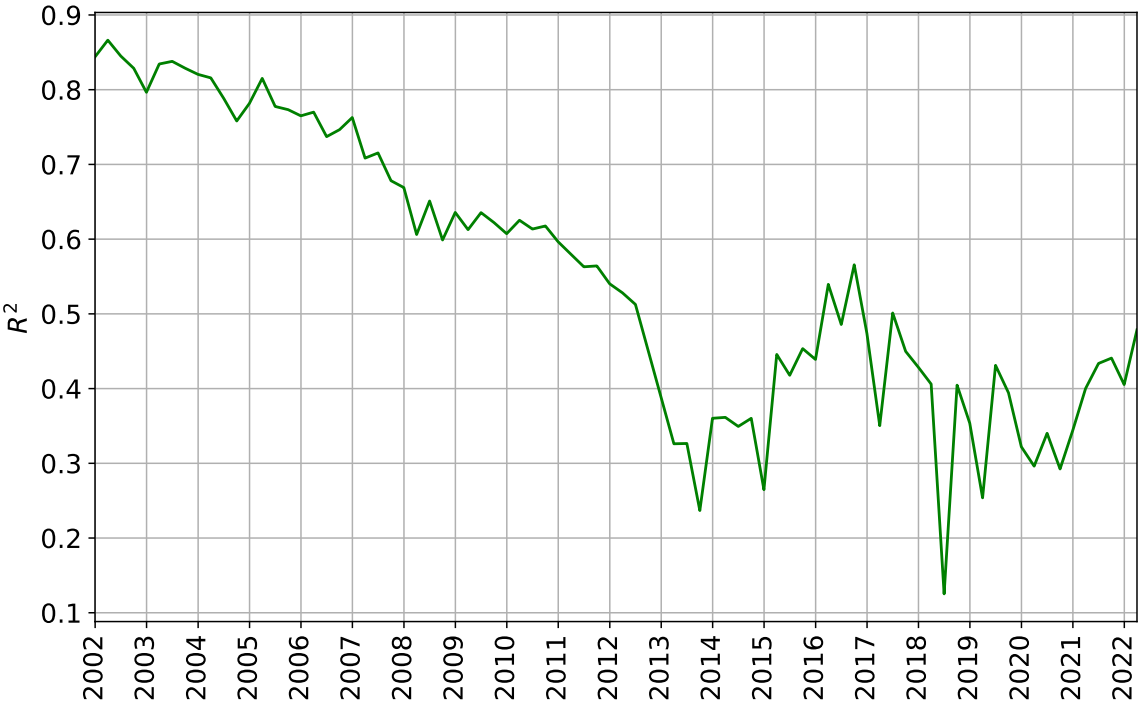
Displays the average variable importance in percent grouped over all investors and dates. Acronyms are as in Chen & Zimmermann (2022). Characteristics from the original Koijen & Yogo (2019) are displayed in orange.

Figure 21: R^2 from OLS Estimation



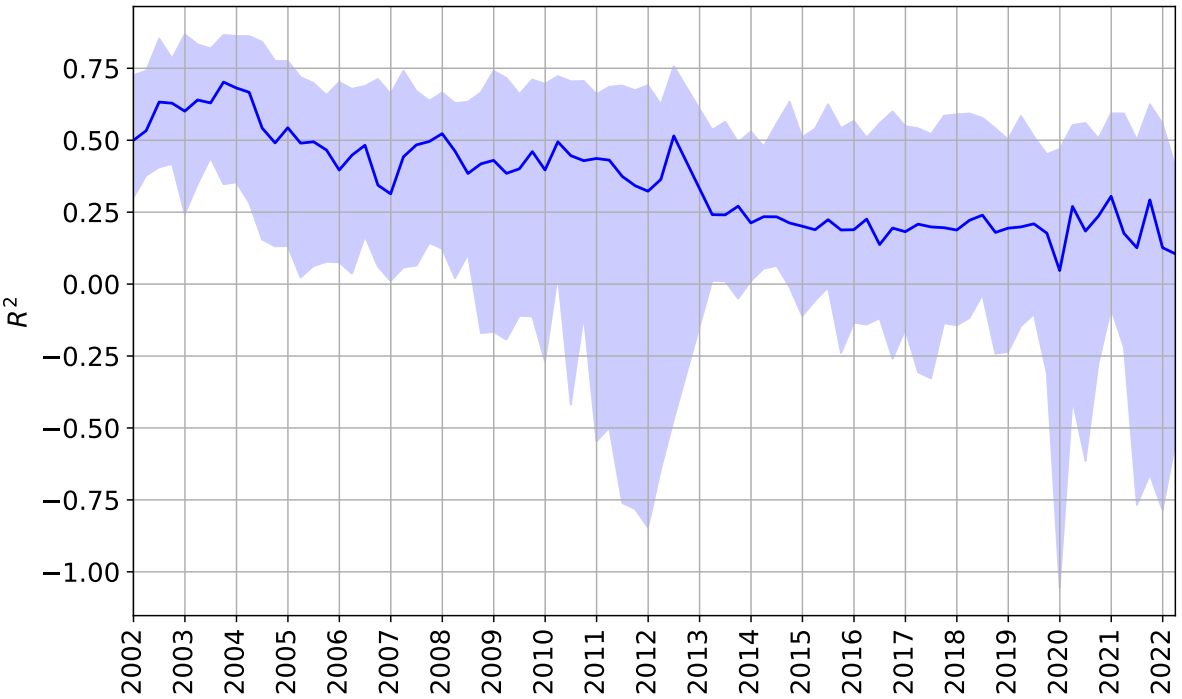
Displays the R^2 for the OLS estimation. OLS estimates are not restricted. Baseline characteristics are used. R^2 is computed separately for each bin in the cross-section. Solid line is the cross-sectional *mean*, shaded area indicates the cross-sectional IQR. Upper Figure only shows R^2 for $log(\delta)$ splitted in groups. Lower Figure shows R^2 for $log(\delta)$ and δ (level) for the entire cross-section.

Figure 22: R^2 Households - GMM Estimation



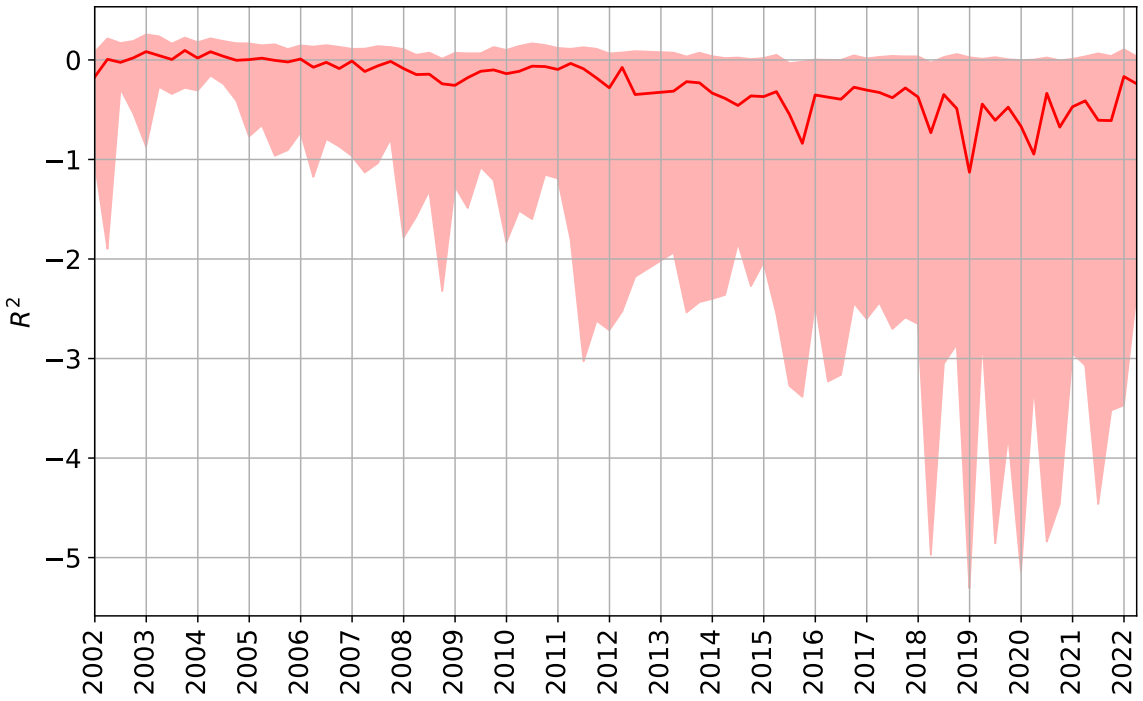
Displays the R^2 from the GMM estimation in each quarter. Baseline characteristics are used.

Figure 23: R^2 Individual Investors - GMM Estimation



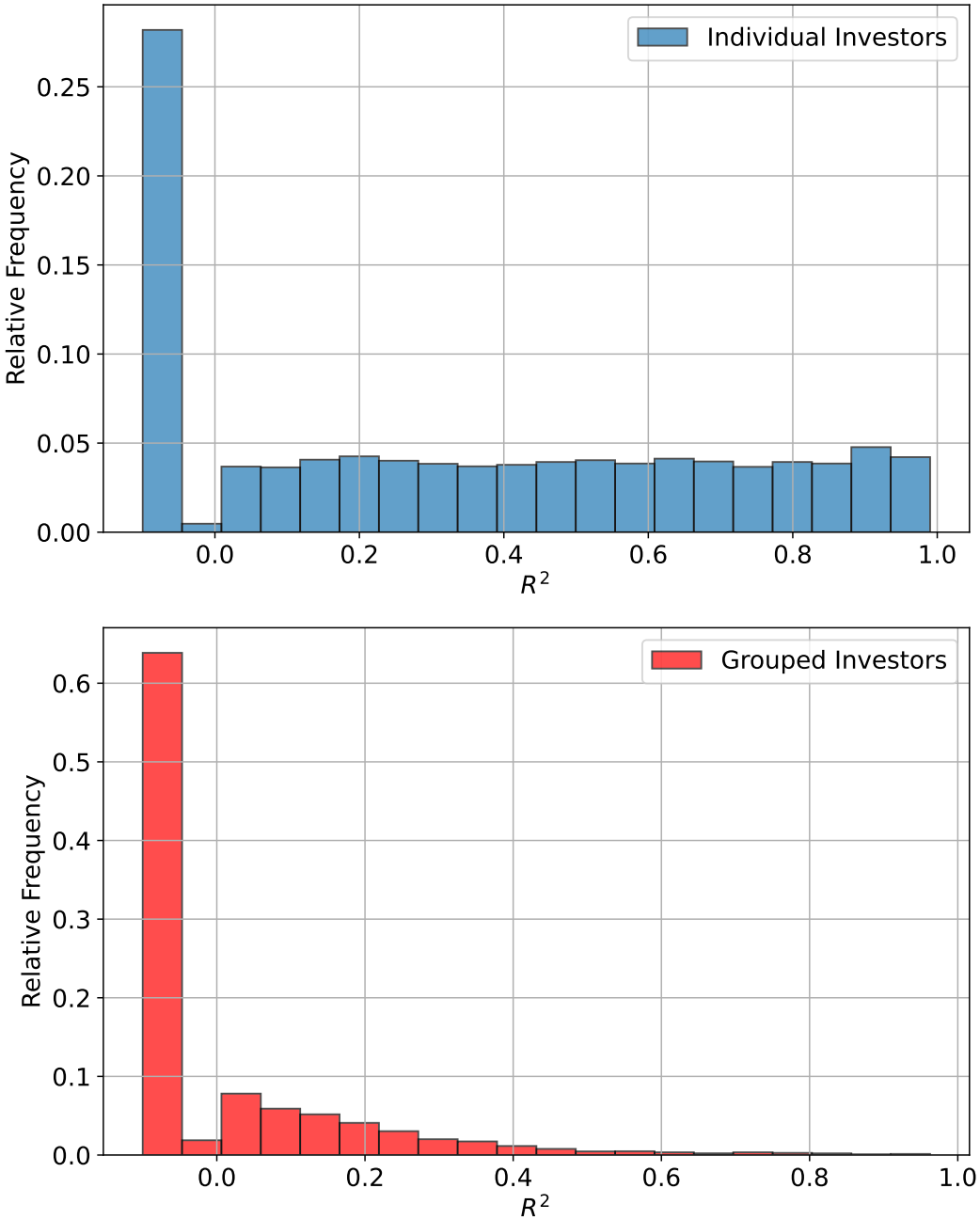
Displays the R^2 from the GMM estimation. Baseline characteristics are used. R^2 is computed separately for each bin in the cross-section. Solid line is the cross-sectional *median*, shaded area is the cross-sectional IQR.

Figure 24: R^2 Grouped Investors - GMM Estimation



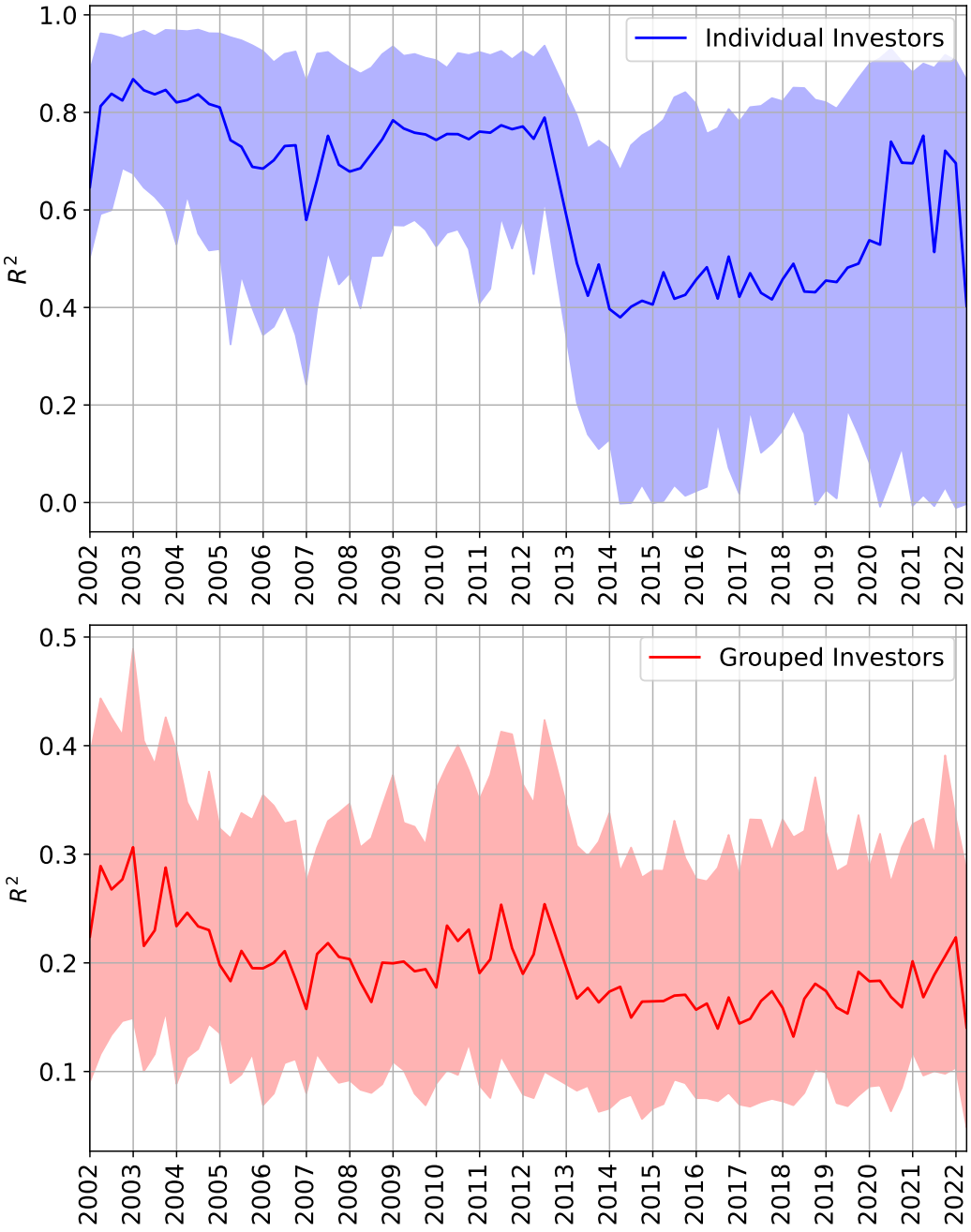
Displays the R^2 from the GMM estimation. Baseline characteristics are used. R^2 is computed separately for each bin in the cross-section. Solid line is the cross-sectional *median*, shaded area is the cross-sectional IQR.

Figure 25: Histogram R^2 - GMM Estimation



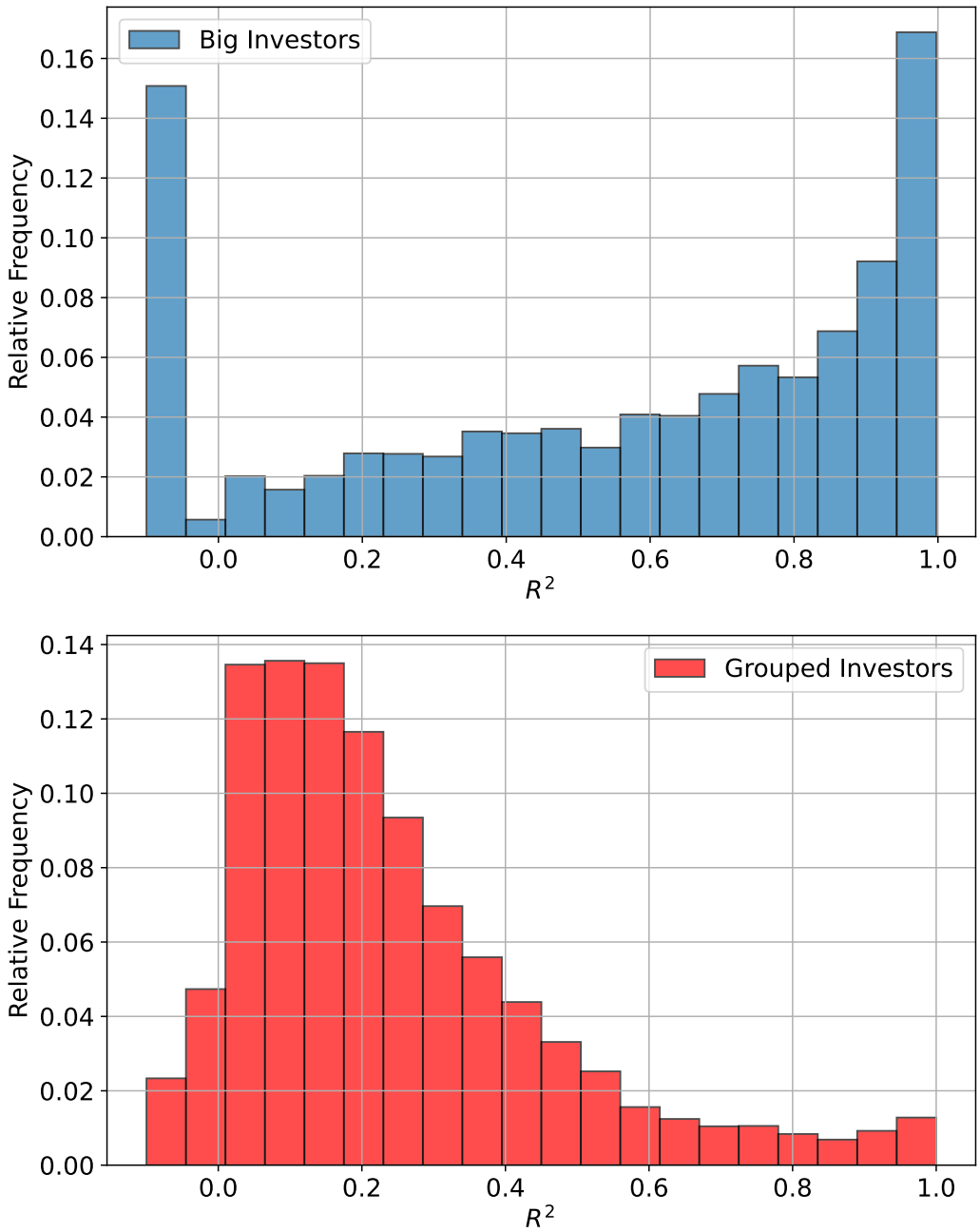
Displays the R^2 from the GMM estimation. Baseline characteristics are used. R^2 is computed separately for each bin in the cross-section. Histogram is built across the entire time series. All negative R^2 are pooled into one bar.

Figure 26: R^2 - NLLS Estimation



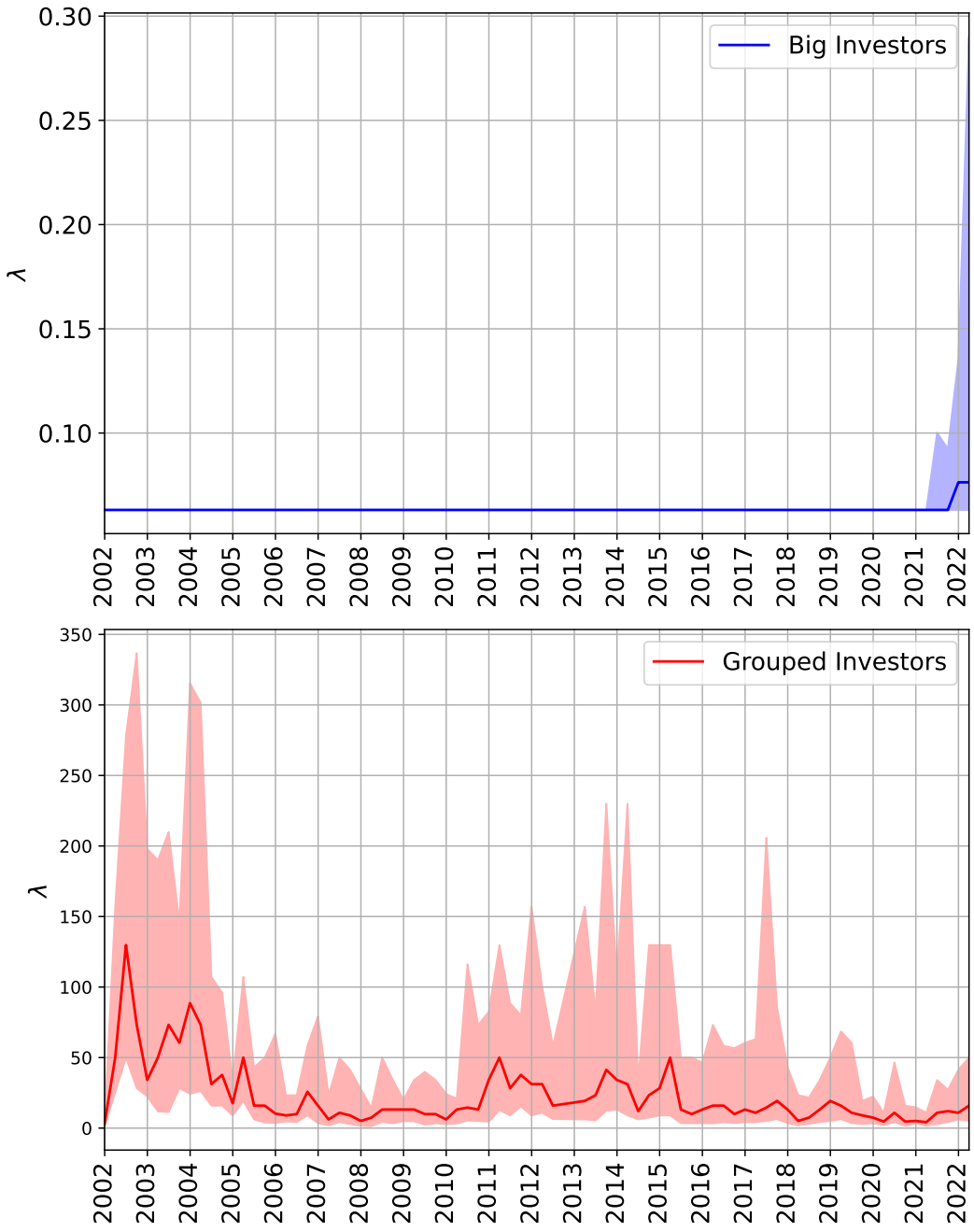
Displays the R^2 from the NLLS estimation with 5-fold cross validation. Baseline characteristics are used. R^2 is computed separately for each bin in the cross-section. Solid line is the cross-sectional *median*, shaded area is the cross-sectional IQR.

Figure 27: Histogram R^2 - NLLS Estimation



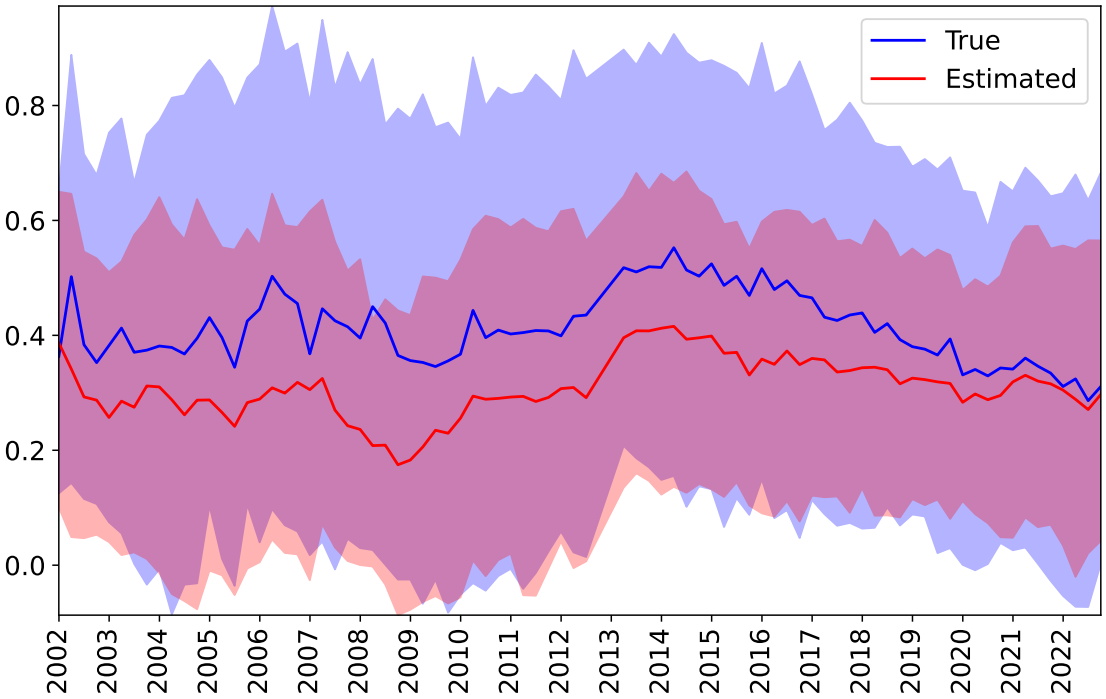
Displays the R^2 from the NLLS with 5-fold cross validation. Baseline characteristics are used. R^2 is computed separately for each bin in the cross-section. Histogram is built across the entire time series. All negative R^2 are pooled into one bar.

Figure 28: Regularization parameter from NLLS Estimation



Displays the regularization parameter from the NLLS estimation with 5-fold cross validation. λ is computed separately for each bin in the cross-section. Solid line is the cross-sectional *median*, shaded area is the cross-sectional IQR.

Figure 29: Comparison Demand Elasticity



Displays the cross-sectional mean of the estimated and true demand elasticity. Shaded area is the IQR.

A Appendix: Derivation

A.1 Demand Elasticity

Derivation of Demand Elasticity (2.5). Fix an investor i and his portfolio $\mathcal{P}_{i,t}$ such that all weights are positive.

Taking the log in (2.3) results in

$$log(\boldsymbol{q}) = -\boldsymbol{p} + log(A) + log(\boldsymbol{w})$$

where the subscripts are omitted as an investor and a time point are fixed. Moreover, the scalar $log(A)$ is added to each element in the vectors. The Jacobian is given by

$$-\frac{\partial log(\boldsymbol{q})}{\partial \boldsymbol{p}} = -\begin{pmatrix} \frac{\partial log(q_1)}{\partial p_1} & \cdots & \frac{\partial log(q_1)}{\partial p_N} \\ \vdots & \ddots & \vdots \\ \frac{\partial log(q_N)}{\partial p_1} & \cdots & \frac{\partial log(q_N)}{\partial p_N} \end{pmatrix}$$

Using the previous, we can quickly obtain

$$-\frac{\partial log(\boldsymbol{q})}{\partial \boldsymbol{p}} = I_N - \begin{pmatrix} \frac{1}{w_1}\frac{\partial w_1}{\partial p_1} & \cdots & \frac{1}{w_1}\frac{\partial w_1}{\partial p_N} \\ \vdots & \ddots & \vdots \\ \frac{1}{w_N}\frac{\partial w_N}{\partial p_1} & \cdots & \frac{1}{w_N}\frac{\partial w_N}{\partial p_N} \end{pmatrix} = I_N - diag(\boldsymbol{w})^{-1} \begin{pmatrix} \frac{\partial w_1}{\partial p_1} & \cdots & \frac{\partial w_1}{\partial p_N} \\ \vdots & \ddots & \vdots \\ \frac{\partial w_N}{\partial p_1} & \cdots & \frac{\partial w_N}{\partial p_N} \end{pmatrix}. \tag{A.1}$$

Thus, what remains to be computed is $\frac{\partial w_j}{\partial p_k}$.

Recall from (2.1) that

$$w_j = \frac{\delta_j}{1 + \sum \delta_n} \qquad \text{with } \delta_j = f\big(\beta_1 p_j + \beta_1 s_j + \textstyle\sum_{k \in K-1} \beta_k x_k\big) \equiv g\big(j\big),$$

where $f(\cdot)$ is the exponential function, $p_j + s_j$ is log market equity, x_k are the other $K-1$ characteristics (including the constant).

Thus, to compute $\frac{\partial w_j}{\partial p_k}$ we simply require

$$\frac{\partial \delta_j}{\partial p_k} = \begin{cases} g'(j) \cdot \beta_1, j = k \\ 0, \text{else} \end{cases}$$

so that

$$\frac{\partial w_j}{\partial p_k} = \begin{cases} \dfrac{g'(j) \cdot \beta_1}{[1 + \sum \delta_n]} - \delta_j [1 + \sum \delta_n]^{-2} \cdot g'(j) \cdot \beta_1 \ , k = j \\ \\ -\delta_j [1 + \sum \delta_n]^{-2} \cdot g'(k) \cdot \beta_1 \qquad\qquad , k \neq j \end{cases} \tag{A.2}$$

Here it becomes apparent why it is tremendously helpful that g is the exponential function. Because plugging (A.2) into (A.1) will not yield a simple way forward to find a restriction on β_1 that ensures that the diagonal elements of (A.1) are strictly positive.

However, since $g'(j) = \delta_j$, we have that

$$\frac{\partial w_j}{\partial p_k} = \begin{cases} w_j\beta_1 - w_j^2 \cdot \beta_1 \ , k = j \\ \\ -w_j w_k \cdot \beta_1 \qquad\qquad\quad , k \neq j \end{cases} \tag{A.3}$$

Plugging (A.3) into (A.1) yields

$$-\frac{\partial log(\boldsymbol{q})}{\partial \boldsymbol{p}} = I_N - diag(\boldsymbol{w})^{-1}\beta_1\Big[diag(\boldsymbol{w}) - \boldsymbol{w}\boldsymbol{w}'\Big]$$

Then, the j-th diagonal element is given by

$$\left(-\frac{\partial log(\boldsymbol{q})}{\partial \boldsymbol{p}}\right)_{[jj]} = 1 - \frac{1}{w_j}\beta_1\Big(w_j - w_j^2\Big) = 1 - \beta_1\Big(1 - w_j\Big).$$

As each $w_j \in [0, 1]$ it holds that the diagonal elements are strictly positive if $\beta_1 < 1$.

A.2 Root-Finding Problem (4.8)

Vectorisation of Objective Function. Firstly, the dataframe for each bin at each point in time is given by the following matrix

$$\begin{pmatrix} y_1 & \boldsymbol{z}_1' & \boldsymbol{x}_1' & \widetilde{x}_1 \\ \vdots & \vdots & \vdots & \vdots \\ y_N & \boldsymbol{z}_N' & \boldsymbol{x}_N' & \widetilde{x}_N \end{pmatrix}, \tag{A.4}$$

where $y_n \geq 0$ is a relative portfolio weight of an investor for a stock, $\boldsymbol{z}_n$ and $\boldsymbol{x}_n$ are $K \times 1$ dimensional vectors and $\widetilde{x}_n$ is the scalar capturing the investor fixed-effect.

From the dataframe (A.4) it is easy to instantly extract
$\boldsymbol{y} = \big(y_1, ..., y_N\big)'$, $\boldsymbol{Z} = \big[\boldsymbol{z}_1, .., \boldsymbol{z_N}\big]'$, $\boldsymbol{X} = \big[\boldsymbol{x}_1, ..., \boldsymbol{x}_N\big]'$ and $\widetilde{\boldsymbol{x}} = \big(\widetilde{x}_1, \ldots, \widetilde{x}_N\big)'$.
We now have all objects to compute the objective function.

1. Compute the vector of errors by

$$\widetilde{\boldsymbol{e}} := \boldsymbol{y} \odot exp\big\{ - \boldsymbol{X}\boldsymbol{\beta} - \widetilde{\boldsymbol{x}}\big\} - \boldsymbol{1}, \tag{A.5}$$

   where $\odot$ denotes the Hadamard product (element-wise multiplication) and the exponential function is applied element-wise which are both implemented efficiently in any numerical library.

2. Compute the vectors $\{\boldsymbol{g}_n(\boldsymbol{\beta})\}_{n=1}^N = \{\boldsymbol{z}_n\widetilde{e}_n\}_{n=1}^N$ by

$$\boldsymbol{G} := \left[\boldsymbol{Z}' \odot \widetilde{\boldsymbol{e}}\right] = \left[\boldsymbol{z}_1\widetilde{e}_1, ..., \boldsymbol{z}_N\widetilde{e}_N\right],$$

so that the n-th column of $\boldsymbol{G}$ corresponds to $\boldsymbol{g}_n$. In Python, computing $\boldsymbol{G}$ is efficiently implemented by `np.newaxis`. Alternatively, $\boldsymbol{G}$ can be computed by stacking each component $\widetilde{e}_n$ into a vector K times repeatedly so that simple matrix multiplication can be used to compute $\boldsymbol{G}$.

3. Lastly, dividing the column sum of $\boldsymbol{G}$ by N yields the objective function $\overline{\boldsymbol{g}}(\boldsymbol{\beta})$. Once again, any numerical library can efficiently compute the column sum.

Jacobian of Objective Function (4.8). We take a bottom-up approach and consider the i-th element of the vector $\boldsymbol{g}_n$ and take the derivative with respect to the j-th element in $\boldsymbol{\beta}$ which is given by

$$\begin{aligned}\frac{\partial g_n(i)}{\partial \beta_j} = z_n(i)\frac{\partial \widetilde{e}_n}{\partial \beta_j} &= z_n(i)\left[\frac{\partial(y_n exp\{-\boldsymbol{x}_n'\boldsymbol{\beta} - \widetilde{x}_n\} - 1)}{\partial \beta_j}\right] \\ &= z_n(i)y_n exp\{-\boldsymbol{x}_n'\boldsymbol{\beta} - \widetilde{x}_n\}[-x_n(j)] = -z_n(i)[\widetilde{e}_n + 1]x_n(j).\end{aligned}$$

Using this, for a single observation the Jacobian is given by

$$\begin{aligned}\frac{\partial \boldsymbol{g}_n}{\partial \boldsymbol{\beta}} &= -\begin{pmatrix} z_n(1)[\widetilde{e}_n + 1]x_n(1) & z_n(1)[\widetilde{e}_n + 1]x_n(2) & \dots & z_n(1)[\widetilde{e}_n + 1]x_n(K) \\ z_n(2)[\widetilde{e}_n + 1]x_n(1) & z_n(2)[\widetilde{e}_n + 1]x_n(2) & \dots & z_n(2)[\widetilde{e}_n + 1]x_n(K) \\ \vdots & \vdots & \ddots & \vdots \\ z_n(K)[\widetilde{e}_n + 1]x_n(1) & \dots & \dots & z_n(K)[\widetilde{e}_n + 1]x_n(K) \end{pmatrix} \\ &= -\boldsymbol{z}_n\boldsymbol{x}_n'[\widetilde{e}_n + 1].\end{aligned}$$

Since derivatives are linear operators it then immediately follows that

$$\frac{\partial \overline{\boldsymbol{g}}}{\partial \boldsymbol{\beta}} = -\frac{1}{N}\sum_{n=1}^N \boldsymbol{z}_n\boldsymbol{x}_n'[\widetilde{e}_n + 1] = -\frac{1}{N}\boldsymbol{Z}'\widetilde{\boldsymbol{X}} \qquad \text{with } \widetilde{\boldsymbol{X}} = \begin{pmatrix} \boldsymbol{x}_1'[\widetilde{e}_1 + 1] \\ \vdots \\ \boldsymbol{x}_N'[\widetilde{e}_N + 1] \end{pmatrix},$$

where $\widetilde{\boldsymbol{X}}$ can once again be efficiently computed with Python's `np.newaxis` and $\widetilde{\boldsymbol{e}}$ was already efficiently computed when constructing the objective function.

A.3 Minimization Problem (4.7)

Standardization. We once again consider the dataframe (A.4). For the computation of $\widetilde{\boldsymbol{e}}$ as in (A.5) we implement the following linear transformation

$$\begin{aligned} exp\left\{\beta(K)\cdot 1+\sum_{k=1}^{K-1}x_n(k)\beta(k)+\widetilde{x}_n\right\} &= exp\left\{\beta(K)+\sum_{k=1}^{K-1}\left(\frac{x_n(k)-\mu_k+\mu_k}{\sigma_k}\right)\sigma_k\beta(k)+\widetilde{x}_n\right\} \\ &= exp\left\{\beta(K)+\sum_{k=1}^{K-1}\mu_k\beta(k)\ +\sum_{k=2}^{K}\left(\frac{x_n(k)-\mu_k}{\sigma_k}\right)\sigma_k\beta(k)+\widetilde{x}_n\right\} \\ &\equiv exp\left\{\gamma(K)+\sum_{k=1}^{K-1}\left(\frac{x_n(k)-\mu_k}{\sigma_k}\right)\gamma(k)+\widetilde{x}_n\right\}, \end{aligned} \tag{A.6}$$

where μ_k is the mean and σ_k is the variance of the k-th variable in $\boldsymbol{x}$ which are computed using dataframe (A.4). Luckily, as $\widetilde{x}_n$ has no coefficient attached to it, it needn't be transformed in any way for stability gains.

Naturally, $\mu_k \in \mathbb{R}$ and $\sigma_k \neq 0$ can be any numbers. For example, this configuration also allows to squeeze the variables into $[0,1]$. We have tried several linear transformations and have found that standardization performs the best in terms of speed and robustness. This is due to the fact that all variables are on the same "scale". If this isn't the case, then for variables with large values such as market equity the Jacobian is much more sensitive towards changes in the respective parameter while for small variable the Jacobian is too insensitive. These properties then also adversely impact the Hessian.

Reducing the Step Size. Instead of (A.6), we additionally include a step size and thus ultimately for the error computation consider the term

$$exp\left\{\gamma(K)\cdot \texttt{step\_size}+\sum_{k=1}^{K-1}\left(\frac{x(k)-\mu_k}{\sigma_k}\right)\gamma(k)\cdot \texttt{step\_size}+\widetilde{x}_n\right\},$$

where we found that a value of 0.01 for the step size leads to major stability increases without slowing down the runtime. It is indeed necessary to scale down the constant as well since the constant was also a source of overflow errors.

The BFGS algorithm will optimize for the parameters $\{\gamma(k)\}_{k=1}^{K}$. Thus, if the algorithm in some step increases a $\gamma(k)$ by 1 unit, it will be effectively only increased by 1 times the step size, i.e. only by 0.01 as per our calibration. This further decreases the overflow errors.

Re-transforming the Estimates. The BFGS-algorithm only outputs $\{\gamma(k)\}_{k=1}^{K}$, but we can easily map them back to $\boldsymbol{\beta}$ by the following

$$\beta(k)=\frac{\gamma(k)\cdot \texttt{step\_size}}{\sigma_k}\quad \forall 1\leq k<K \quad \text{and} \quad \beta(K)=\gamma(K)\cdot \texttt{step\_size}-\sum_{k=1}^{K-1}\mu_k\beta(k).$$

Therefore, the restriction (2.7) can be easily implemented by considering

$$\beta(1) \leq 0.99 \Rightarrow \gamma(1) \leq \frac{0.99\sigma_1}{\texttt{step\_size}},$$

which is the bound we feed into the L-BFGS-B solver for the first argument. Thus, the linear transformation is completely non-invasive.

Iterative BFGS Application with Weighting Matrix. The most crucial aspect to achieve convergence is the iterative application of the BFGS solver. We proceeded in the following manner. Initialize the tolerance to a high level which we set to 0.001. Additionally, we initialize $\boldsymbol{W}^0 = \boldsymbol{I}$. Then, apply the BFGS solver to minimize the objective Q along with having incorporated the previous linear transformations to obtain the first estimate $\boldsymbol{\gamma}^{(0)}$. As an initial guess, we have found that $\boldsymbol{\gamma} = 0.5/\texttt{step\_size}$ yields good results.

Then, we perform an iterative procedure. In each iteration we scale down the tolerance by a factor of $1,000$ compared to the previous iteration, use $\boldsymbol{\gamma}^{(i)}$ as the new initial guess and update the weighting matrix according to

$$\boldsymbol{W}^{(i+1)} = \left(\frac{1}{N} \sum_{n=1}^{N} \boldsymbol{g}_n(\boldsymbol{\gamma}^{(i)}) \boldsymbol{g}_n(\boldsymbol{\gamma}^{(i)})' \right)^{-1},$$

We stop once the weighting matrix converges[17]. In the end, we re-transform $\boldsymbol{\gamma}$ to $\boldsymbol{\beta}$.

The weighting matrix can be omitted, even though it does increase numerical stability and accuracy, but the step-wise reduction of the tolerance is absolutely crucial.

B Appendix: Data

B.1 Characteristics

Table 7: List of additional characteristics from Chen & Zimmermann (2022).

Name	Acronym
Accruals	
Accruals	Accruals
Abnormal Accruals	AbnormalAccruals
Percent Operating Accruals	PctAcc
Percent Total Accruals	PctTotAcc
Asset Composition	
Cash to assets	Cash
Net Operating Assets	NOA
Composite Accounting	

Continued on next page

[17] When using the restriction (2.7), the algorithm can oscillate so that the weighting matrix never converges. Therefore, we include a stopping condition after a certain number of iterations which are never more than 50.

Table 7 – continued from previous page

Name	Acronym
Excluded Expenses	ExclExp
Earnings Growth	
Earnings Surprise	EarningsSurprise
Earnings streak length	NumEarnIncrease
External Financing	
Convertible debt indicator	ConvDebt
Change in current operating liabilities	DelCOL
Change in financial liabilities	DelFINL
Net debt financing	NetDebtFinance
Net external financing	XFIN
Debt Issuance	DebtIssuance
Info Proxy	
Firm age based on CRSP	FirmAge
Investment	
Asset growth	AssetGrowth
Growth in book equity	ChEQ
Change in equity to assets	DelEqu
Change in long-term investment	DelLTI
Change in net operating assets	dNoa
Change in ppe and inv/assets	InvestPPEInv
Growth in long-term operating assets	GrLTNOA
Inventory Growth	ChInv
Change in Net Noncurrent Op Assets	ChNNCOA
Change in Net Working Capital	ChNWC
Change in current operating assets	DelCOA
Change in net financial assets	DelNetFin
Employment growth	hire
Total accruals	TotalAccruals
Change in capital inv (ind adj)	ChInvIA
Change in capex (two years)	grcapx
Change in capex (three years)	grcapx3y
Leverage	
Book leverage (annual)	BookLeverage
Long Term Reversal	
Long-run reversal	LRreversal
Medium-run reversal	MRreversal
Payout Indicator	
Dividend Initiation	DivInit
Dividend Omission	DivOmit
Profitability	
Cash-based operating profitability	CBOperProf
Gross profits / total assets	GP
Return on assets (qtrly)	roaq
Net income / book equity	RoE
Sales Growth	
Revenue Surprise	RevenueSurprise
Sales growth over overhead growth	GrSaleToGrOverhead
Valuation	
Enterprise component of BM	EBM
Liquidity	
Days with zero trades (6m)	zerotrade
Other	

Continued on next page

Table 7 – continued from previous page

Name	**Acronym**
Change in Taxes	ChTax
Industry concentration (sales)	Herf
Industry concentration (equity)	HerfBE
Off season long-term reversal	MomOffSeason
Off season reversal years 6 to 10	MomOffSeason06YrPlus
Return seasonality years 2 to 5	MomSeason
Return seasonality years 6 to 10	MomSeason06YrPlus
Return seasonality last year	MomSeasonShort
Operating leverage	OPLeverage
Taxable income to income	Tax
Industry concentration (assets)	HerfAsset
Spinoffs	Spinoff