Asset Demand Systems via Data Augmentation: Competition and Differentiation in Asset Management

Yurii Handziuk*
HEC Paris

January 12, 2025

[CLICK HERE FOR THE MOST RECENT VERSION]

Abstract

Many institutional investors hold portfolios with few holdings. This makes it challenging to precisely estimate their individual demand. In this paper, I seek to make two contributions. First, I propose a data augmentation technique based on the generation of data-driven and economically interpretable synthetic assets. I show that this data augmentation acts as an adaptive nonlinear shrinkage which automatically adjusts the shape of the penalty to the cost of overfitting faced by the nonlinear demand function estimator. The resulting estimation technique leads to substantial improvement in cross-out-of-sample R^2 for estimation of both low-dimensional and high-dimensional demand functions. Second, I use the proposed methodology to construct a measure of investor differentiation. Using the Morningstar mutual fund ratings reform in 2002 as a shock to competition for alpha, I show that mutual funds escape the increased competition intensity by differentiating from their competitors.

Keywords: Asset demand system, Asset management, Competition, Differentiation, Machine learning, Data augmentation, Synthetic data

*E-mail: yurii.handziuk@hec.edu. I am very grateful to Augustin Landier, Thierry Foucault, Johan Hombert, and Stefano Lovo for their continuous guidance and support. I am also thankful to Bruno Biais, Jean-Edouard Colliard, Francois Derrien, Tjeerd De Vries, Matthias Efing, Quirin Fleckenstein, Jacques Olivier, Ioanid Rosu, Daniel Schmidt, Irina Zviadadze and seminar participants at HEC Paris, Hi! Paris Center on Artificial Intelligence for Science, Business and Society for very helpful comments and suggestions. I acknowledge the funding support of the Hi! Paris Center on Artificial Intelligence for Science, Business and Society. This work has benefited from a State grant managed by the Agence Nationale de la Recherche under the Investissements d'Avenir programme with the reference ANR-18-EURE-0005 / EUR DATA EFM.

1 Introduction

Many institutional investors hold portfolios with few holdings. Due to the scarcity of portfolio holdings data,[1] it is challenging to obtain precise estimates of the *individual* demand of institutional investors as a multivariate function of the set of stock characteristics (Koijen and Yogo (2019), Koijen et al. (2023)).[2] This issue is further exacerbated if one's goal is to estimate complex, high-dimensional demand functions in line with the modern asset pricing literature.[3] This limits the pursuit of research questions requiring highly flexible measurement of institutional investors' portfolio decisions, such as studies on differentiation and innovation in terms of investment strategies.

In this paper, I seek to make two contributions. First, to address the scarcity of holdings data, I propose a data augmentation technique based on the generation of data-driven and economically interpretable synthetic assets. I show that augmentation of institutional investor holdings with the proposed synthetic assets acts as an adaptive shrinkage estimator. The rate of shrinkage imposed on the coefficients by synthetic assets adapts to the functional form of the demand function estimator, leading to more efficient shrinkage. In the case of nonlinear GMM estimation in Koijen and Yogo (2019) and Koijen et al. (2023), who model asset demand as an exponential function of stock characteristics, the coefficients are shrunk according to an exponential penalty on their deviation from a shrinkage target.

[1] For instance, the median number of holdings available for an econometrician to study an individual active equity mutual fund is just around 70 holdings per quarter (in a merged Thomson Reuters s12-CRSP Mutual Fund dataset). For 13F institutions (Thomson Reuters s34), the median number of holdings is around 100 holdings per quarter. For descriptive statistics on the number of holdings in active mutual funds and 13F institutions, see Tables A.1 and A.2, respectively. The data scarcity is even more severe in bond holdings data: see Nenova (2024), for example.

[2] Koijen and Yogo (2019) note: "We estimate the coefficients by institution whenever there are more than 1,000 strictly positive holdings in the cross section. For institutions with fewer than 1,000 holdings, we pool them with similar institutions in order to estimate their coefficients... While the cutoff of 1,000 is arbitrary, a lower cutoff of 500 causes convergence problems for our estimator in some cases. We set the total number of groups at each date to target 2,000 strictly positive holdings ... per group." Also, Koijen and Yogo (2019) model investors' demand as a function of a small set of asset characteristics: "Our specification is based on a parsimonious and relevant set of characteristics for explaining expected returns and factor loadings... We are concerned about collinearity between characteristics and overfitting if we consider a larger model with more characteristics."

[3] See, for example, Kelly et al. (2024), Kelly et al. (2022), Didisheim et al. (2023), Kozak et al. (2020), Gu et al. (2020), Martin and Nagel (2022) for high-dimensionality in asset pricing and see Kaniel et al. (2023), Gabaix et al. (2024) for examples in asset management literature. For adoption of AI, machine learning, and alternative data by modern asset managers, see Abis (2020), Bonelli and Foucault (2023), Dugast and Foucault (2023), Bonelli (2022).

The intuition of the mechanism through which the data augmentation with synthetic assets acts as a shrinkage is depicted in Figure 2. When the model of interest is very complex relative to the number of available observations, the model overfits the training data, which leads to poor generalization on new, unseen data. The cornerstone of the data augmentation as a shrinkage technique is to initially fit a simplified, robust to overfitting model, and then generate synthetic data points as predictions obtained from this simplified model. Then, the complex model of interest is estimated on the *augmented* dataset consisting of both original and synthetic data points, which shrinks the complex model towards the predictions made by the simple model, reducing the complex model's variance and mitigating overfitting.

In the context of estimating asset demand functions, I propose the following data augmentation algorithm. First, I estimate the log-linearized specification of the demand function of a given institutional investor i at a given point in time t using a penalized linear estimator that is robust to overfitting (such as linear ridge regression). Second, I generate synthetic assets whose characteristics provide a canonical basis for the space of characteristics used in the demand function estimation. Then, the synthetic demand of investor i at time t for each *characteristic-basis* synthetic asset is generated as out-of-sample predictions of the *simplified* model obtained in the first step. In the third and final step, for each separate investor i at a given time t, the *augmented* dataset – which comprises of both original and synthetic assets – is used to estimate the demand function with nonlinear GMM of Koijen and Yogo (2019). This last step can be performed *without* the pooling of holdings across investors, which would have been required in the nonlinear GMM estimation approach of Koijen and Yogo (2019), Koijen et al. (2023). The composition of the augmented dataset is illustrated in Figure 1, Panel A.

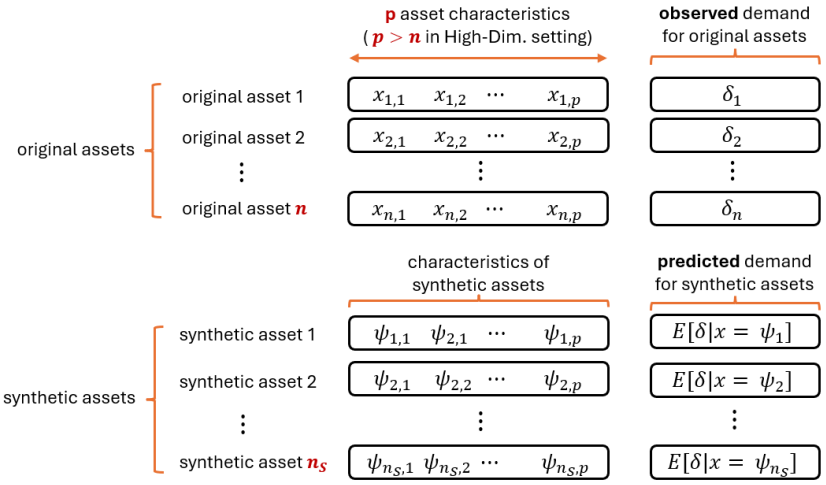
The importance of my methodological contribution is twofold. First, my approach provides a better explanation of the asset demand of institutional investors. As illustrated in Panel B of Figure 1, the data augmentation approximately *doubles* the mean cross-out-of-sample R^2 for estimating standard low-dimensional demand functions compared to the state-of-the-art method of Koijen et al. (2023). When the demand function is extended to the high-dimensional set of stock characteristics comprising

characteristics from factor zoo and industry dummies, the mean cross-out-of-sample R^2 increases by a factor of about 2.5. Second, from an economic standpoint, the capability to estimate high-dimensional asset demand functions individually for each investor is central to providing a demand-function analogue to measures of investor differentiation used in the literature on industrial organization of asset management. In this way, I provide a link between the literature on demand-based asset pricing (Koijen and Yogo (2019), Koijen et al. (2023)) and the literature on industrial organization of asset management.

Unlike recent studies addressing the issue of missing data in finance panels (notably, Giglio et al. (2021), Freyberger et al. (2024), Bryzgalova et al. (2024), Kaniel et al. (2023)), I propose a data augmentation technique that serves as a regularization method, whereby synthetic assets effectively prevent overfitting and issues arising from multicollinearity. Imputation of missing values is fundamentally distinct from generating fully-synthetic data points for regularization. In the former, an econometrician uses conditional mean expectations of regressors with missing values *given* non-missing values of other regressors. In other words, imputation of some missing regressors for a given observation j requires to observe at least some regressors for the exact same observation j. In my approach of data augmentation with characteristic-basis synthetic assets, the values of regressors are fully synthetic and are selected flexibly using cross-validation allowing one to create additional data points even if *all* regressors are missing.

In my second contribution, I propose a multivariate, well-suited for modern high-dimensional settings measure of similarity between the investment strategies pursued by institutional investors. The proposed measure – which I refer to as *asset demand function similarity* (ADFS) – is based on the cosine similarity between the vectors of the estimated asset demand function loadings of investors. Unlike prospectus-based measures, my measure captures which investment strategies are pursued by mutual funds, rather than how mutual funds advertise themselves. Further, unlike univariate portfolio-weighted characteristic-based measures, my measure captures the correlations among the asset characteristics. Last but not least, since the measure is based on demand functions in a structural model, it can directly map the differentiation and innovation of institutional investors into asset prices through the approach of Koijen

Panel A: Data augmentation of investor's holdings



Panel B: Performance of demand function estimators across fund-years

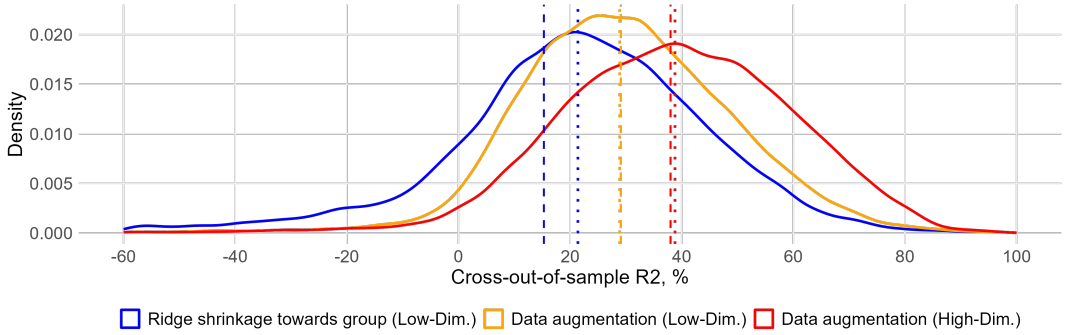


Figure 1: Data augmentation: main result

Note: Panel A illustrates the data augmentation approach. Suppose that for a *given* investor i at time t we observe n original assets with characteristics $x_i \in \mathbb{R}^{p\times 1}$, $i \in \{1,,...,n\}$ and investor i's time-t demand for those assets is δ_i, $i \in \{1,,...,n\}$ (the subscript t is omitted for brevity). Then, the original observations of portfolio holdings are augmented by adding n_S synthetic assets with characteristics $\psi_i \in \mathbb{R}^{p\times 1}$, $i \in \{1,,...,n_S\}$. The *synthetic* demand for synthetic assets is generated as a prediction from *simplified* model. Finally, the model of interest is estimated using both n original and n_S synthetic assets. Panel B shows the distribution of cross-out-of-sample R^2 across fund-years. Cross-out-of-sample R^2 measures how well $1/5$ of fund i's holdings at time t are explained by a demand function estimated (including hyper-parameter selection) on the remaining $4/5$. To construct R^2, this process is repeated across all 5 folds. Vertical dashed (dotted) lines show the mean (median) of the distribution of cross-out-of-sample R^2 for each estimation approach. Label "Ridge shrinkage towards group" corresponds to the approach proposed in Koijen et al. (2023). "Data augmentation" refers to the estimation of asset demand functions on the *augmented* dataset using nonlinear Generalized Method of Moments (GMM) of Koijen et al. (2023) *without* shrinkage towards the group priors. For low-dimensional demand functions (labelled as "Low-Dim."), the set of asset characteristics comprise of the superset of baseline characteristics used in Koijen and Yogo (2019), Koijen et al. (2023), resulting in $p = 8$ characteristics. For high-dimensional demand functions (labelled as "High-Dim."), the baseline set of characteristics is extended by 77 stock characteristics from "factor zoo" and 85 SIC 2-digit industry dummies, resulting in a total of $p = 170$ characteristics. The figure shows only the low-dimensional version of the nonlinear GMM with shrinkage towards group target since for high-dimensional demand functions, the estimator systematically does not converge. Following Koijen et al. (2023), the demand function is estimated annually for a given fund. The sample consists of active equity mutual funds in the U.S. between 1992 and 2022. To mitigate the impact of outliers on the mean, I winsorize the distribution of R^2 at 2.5%.

and Yogo (2019), Koijen et al. (2023).

Using the proposed measure, I show that mutual funds escape the increased competition through the differentiation from competing funds in the same style. To address the endogeneity of competition and differentiation, I use the Morningstar mutual fund ratings reform in 2002 – introduced by Ben-David et al. (2020) – as a plausibly exogenous shock to the style-level competition for alpha. Before June 2002, Morningstar used to assign ratings based on the ranking of *all* U.S. equity mutual funds without taking into account the style of each specific fund. In June 2002, Morningstar changed drastically their approach by starting to assign ratings based on *within-style* rankings of funds. In this way, funds that belonged to poor-performing styles obtained, on average, a positive shock to their Morningstar rating, while funds in well-performing styles received, on average, a negative shock to their rating (Ben-David et al. (2020)). In the 2SLS estimation of the impact of increased fund flows into styles on the similarity of mutual funds' asset demand functions, the relevance condition is satisfied since Morningstar ratings attract flows, as shown by a strong first stage in my analysis. The exclusion restriction relies on the identifying assumption that the only channel through which the Morningstar mutual fund ratings reform affected mutual funds' portfolio decisions is through style-level flows.

My finding provides plausibly causal evidence of a new channel of mutual fund differentiation in the space of investment strategies, complementing recent studies on mutual fund differentiation and innovation.[4] Using textual analysis of mutual fund prospectuses, Kostovetsky and Warner (2020) show that small and young fund families offer more unique mutual funds, and that more unique funds have lower flow-performance sensitivity. Based on a measure of uniqueness of mutual fund prospectuses, Bonelli et al. (2021) show that after receiving a negative signal about their quality, poorly performing mutual funds differentiate on non-performance dimension by offering more unique products catering to niche clientele. Abis and Lines (2024) find that capital flows to mutual funds respond negatively when mutual funds deviate from their prospectus-based strategy peer groups.

[4]See Kostovetsky and Warner (2020), Abis and Lines (2024), Bonelli et al. (2021), Lettau et al. (2018).

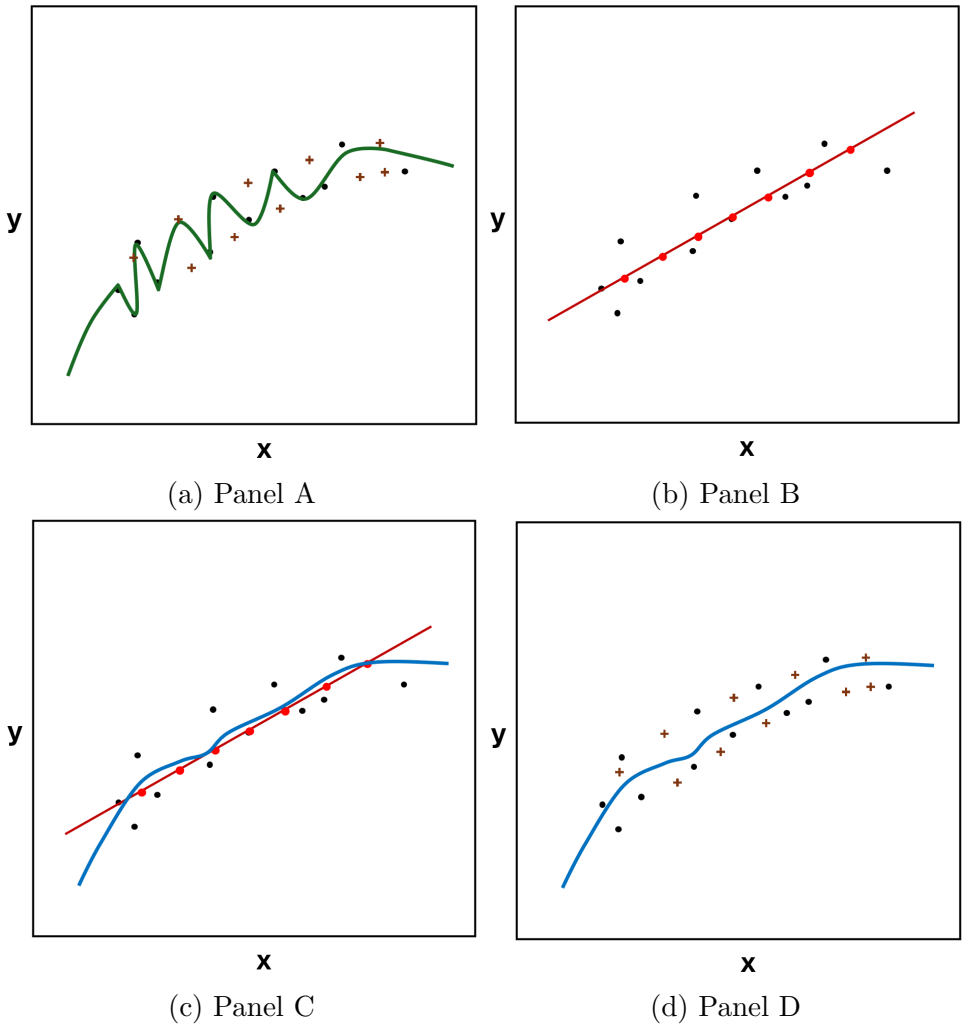
In my paper, I provide evidence that after the increase in competition for alpha due to the increased capital flows to competitors, mutual funds differentiate from their style-based peers in terms of investment strategies, measured as the cosine similarity between their asset demand functions. The finding is robust to controlling for the mutual fund performance and the performance of the funds' peers, suggesting that my finding is not driven by the quality of the fund or its competitors, but rather by the increased difficulty of the search for alpha in the presence of larger diseconomies of scale at the style level. Using *partial* cosine similarity between the demand function loadings on a specific subset of stock characteristics, I find that active mutual funds mostly differentiate across the stock characteristics representing investment, accruals, and profitability, but not through the differentiated industry exposures.

My results suggest that policies aimed at increasing capital flows into the mutual fund sector (such as decreasing the costs associated with investing in mutual funds or increasing tax incentives) can lead to more innovation by mutual funds. Given the finding that mutual funds dynamically differentiate across investment strategies other than size and value, my paper also provides further evidence in support of customized-peer mutual fund performance evaluation (Hoberg et al. (2018), Abis and Lines (2024)) and fund-specific estimation of the diseconomies of scale (Berk and Green (2004), Barras et al. (2022)).

1.1 Related Literature

My paper contributes to three strands of literatures: literature on the estimation of investor demand functions, literature on industrial organization of asset management industry, and the literature that employs machine learning to study complex, high-dimensional phenomena in financial economics.

Estimation of investor demand functions. Modern empirical asset pricing literature is increasingly using the demand-based asset pricing approach of Koijen and Yogo (2019), Koijen et al. (2023) to study the asset pricing implications of the portfolio choices of institutional investors. Examples of such studies include Koijen and Yogo (2024), Koijen et al. (2021), van der Beck (2022), van der Beck (2021), Haddad et al. (2021), Huebner (2023), Bretscher et al. (2022), Plazzi et al. (2023), Noh et al. (2020).



(a) Panel A (b) Panel B

(c) Panel C (d) Panel D

Figure 2: Data augmentation as a shrinkage method: Stylized illustration

Note: Panel A illustrates how highly-flexible and complex function $y = f(x)$ (green solid curve) overfits the in-sample data (black dots), and thus, generalizes poorly to the out-of-sample data (denoted by brown crosses). In Panel B, a simplified and robust to overfitting function $y = g(x)$ (solid red line) is fitted to the in-sample data only. Then, in Panel C, the highly-flexible, complex function $f()$ (blue solid line) is fitted on the *augmented dataset* that contains both the original in-sample data (black dots) and synthetic data points (red dots) generated as a prediction from the simplified model. Panel D shows how the complex function $f()$ fitted with data augmentation provides a better description of the underlying true data generating process and generalizes well on the previously unseen data.

I contribute to this literature by proposing a data augmentation-based approach to the estimation of asset demand functions which extends Koijen and Yogo (2019), Koijen et al. (2023) on two key aspects simultaneously: 1) estimation of the time-varying nonlinear demand functions *individually* for each institutional investor without relying on pooling of institutional investors together or shrinkage towards first-step estimates obtained on pooled sample; 2) estimation of high-dimensional demand functions where the number of stock characteristics potentially relevant for investors' demand can be similar to or even exceed the number of original observations of stock holdings. While

Haddad et al. (2021) and Huebner (2023) resort to log-linearization of (low-dimensional) structural equation to obtain investor-specific estimates – which restricts the set of assets used in estimation to those with strictly positive portfolio weights – my approach of data augmentation allows for the estimation nonlinear GMM proposed by Koijen and Yogo (2019), Koijen et al. (2023), which permits zero-weight holdings in estimation. Without zero-weight holdings, the demand function only describes the intensive margin of investor's demand: conditional on choosing to hold asset j, how much of the asset is held? With nonlinear GMM approach of Koijen and Yogo (2019), Koijen et al. (2023), the extensive margin (which assets are held) is incorporated into the demand function estimation as well. While I employ data augmentation approach in the specific case of Koijen and Yogo (2019), Koijen et al. (2023), the augmentation of holdings with synthetic assets proposed in my paper can also be used in future work that applies extremely flexible machine learning techniques to the estimation of investors' demand, for example to learn asset embeddings or investor embeddings in Gabaix et al. (2024).

Competition and differentiation of institutional investors. There has been a recent and rapid growth in the literature studying the industrial organization of asset management, and specifically, the questions regarding competition, differentiation, and innovation among asset managers. Studies in this literature introduced various measures of similarity between the investment vehicles to capture different dimensions of competition and differentiation. Wahal and Wang (2011) use overlap in portfolio holdings as a measure of similarity between the incumbents and new entrants in mutual funds industry. Using return-based factor exposures of funds, Li and Qiu (2014) show that mutual funds with more extreme factor exposures charge higher fees. Kostovetsky and Warner (2020) propose a measure of mutual fund uniqueness based on the textual analysis of mutual fund prospectuses. Abis and Lines (2024) suggest to use the k-means clustering algorithm on the corpora of mutual fund prospectuses to construct prospectus-based strategy peer groups (SPGs). Hoberg et al. (2018) propose to measure the distance between the pair of mutual funds as the distance between the vectors of investment-weighted stock characteristics of each fund. Although their approach does capture the actual investment strategy of funds – rather than what funds advertise in their prospectuses – the measure based on portfolio-weighted stock characteristics is univariate by construction and does not take into account the correlation between

the stock characteristics.[5] I contribute to the literature by proposing a multivariate, suitable for modern high-dimensional settings measure that captures which investment strategies in the space of observable asset characteristics are actually pursued by mutual funds rather than those advertised in prospectuses. Inspired by the success of transformer-based BERT models in natural language processing, Gabaix et al. (2024) propose a machine learning method AssetBERT to learn latent vector representations of assets, called asset embeddings, and analogously defined InvestorBERT to obtain investor embeddings. The authors further suggest that investor embeddings can be used to measure the similarity of investors in the *latent* space of embeddings. There are two key advantages of constructing similarity measures based on the investor-level high-dimensional version of Koijen et al. (2023) compared to investor embeddings proposed in Gabaix et al. (2024). First, the similarity of demand function loadings directly captures the investment strategy pursued by an investor in the space of interpretable stock characteristics, whereas investor embeddings are latent by construction. Second, measuring similarity in terms of high-dimensional version of Koijen et al. (2023) demand functions allows one to explore *which* observable types of stock characteristics (e.g., measures of profitability or industry dummies) institutional investors differentiate along.[6] For example, using *partial* cosine similarity based on the demand function loadings on characteristics from the profitability theme (as defined in Jensen et al. (2023)), one can estimate the extent to which investor differentiation is driven by differences in exposure to stock characteristics measuring profitability.

Machine learning and high-dimensionality in finance. A growing literature in finance adapts machine learning tools to answer questions in financial economics. Machine learning has been applied to various settings in empirical asset pricing: Cong et al. (2021), Kelly et al. (2024), Kelly et al. (2022), Didisheim et al. (2023), Kozak

[5]In a refinement of the baseline method, Hoberg et al. (2018) propose to sequentially orthogonalize stock characteristics with respect to some pre-defined by econometrician rule before constructing fund-level characteristics, yet this approach still does not allow for simultaneous multivariate estimation and is likely to be sensitive to the pre-specified orthogonalization sequence. Further, the noise in sequential orthogonalization accumulates after each orthogonalization, raising concerns about very noisy residual characteristics when the number of stock characteristics grows large.

[6]Under appropriate exclusion restrictions, latent asset characteristics – such as asset embeddings obtained using AssetBERT in Gabaix et al. (2024) – can also be included in the demand function estimation, providing a way to jointly estimate demand function loadings on both observable and latent asset characteristics in the estimation of a high-dimensional demand system. This would allow an empiricist to include additional information about asset types that are not incorporated in the observed asset characteristics.

et al. (2020), Gu et al. (2020), and Martin and Nagel (2022); asset management: Kaniel et al. (2023) Gabaix et al. (2024); and corporate finance: Hommel et al. (2021). To the best of my knowledge, my paper is the first to propose data augmentation as a shrinkage method to regularize the high-dimensional estimation problem and to address the scarcity of data in asset pricing and asset management panel data.

Data augmentation in computer science and statistics. The data augmentation approach to regularization is commonly used in computer science. In the seminal paper on deep learning for image classification, Krizhevsky et al. (2012) augment the original dataset of 1.2 million images by generating label-preserving transformations of the original images to train the neural network architecture with 60 million parameters. The label-preserving transformations – such as horizontal reflection of the image, color distortion or extracting random patches from the original images – do not alter the label of the image (e.g., whether it is a dog or a flower) while creating additional training data that is synthesized according to a pre-determined rule.[7] The most closely related to my data augmentation approach are Li and Liu (2022) and Huang et al. (2020). Li and Liu (2022) propose an adaptive noisy data augmentation approach that implements well-known penalized estimators such as Lasso, Elastic Net, SCAD by adding observations consisting of noise drawn from a distribution specific to the chosen type of penalized estimator. In Huang et al. (2020), the authors generate noisy synthetic data from the predictive distribution of a simpler model to regularize a high-dimensional "working model". The data augmentation approach in my paper follows a similar logic to Huang et al. (2020) in that the more complicated demand function estimator is disciplined by the output obtained from a simpler model. However, unlike Huang et al. (2020) who rely on the addition of noisy data and subsequent convergence of this noise to a penalty, my data augmentation approach is deterministic since the synthetic assets are constructed as a basis spanning the space of asset characteristics. This provides two advantages. First, the synthetic dataset is *parsimonious* in that it does not require a large number of synthesized observations to obtain precise regularization, which might be very computationally costly. Second, the deterministic nature of the synthetic as-

[7]The label-preserving transformations of the original data are clearly interdependent with original data. However, Krizhevsky et al. (2012) note that the benefit from such data augmentation is substantial "*...The resulting training examples are, of course, highly interdependent. Without this scheme, our network suffers from substantial overfitting, which would have forced us to use much smaller networks.*"

sets ensures that the estimates of demand function loadings are not dependent on the particular sequence of randomly generated noise, which fosters replicability of results across researchers. My paper is also related to the computer science literature on model collapse due to an excess of synthetic data in training datasets (see, e.g. Dohmatob et al. (2024)). Papers in this literature focus on settings where the distinction between original and synthetic data is practically impossible,[8] and show that an overabundance of synthetic data leads to the deterioration of the quality of machine learning models. Unlike this literature, in my data augmentation approach, the identity of synthetic data is known by construction since it is generated by the method.

The rest of this paper is organized as follows. Section 2 describes the proposed data augmentation methodology and its application to asset demand system estimation. In section 3, the proposed methodology is tested on simulated investor holdings. Section 4 provides a description of the data. Section 5 presents the cross-out-of-sample validation of the method. Section 6 outlines the identification strategy for the application of the high-dimensional demand functions to study competition-induced differentiation and presents the empirical results. Section 7 shows robustness checks and section 8 concludes.

2 Methodology

2.1 Estimation of Asset Demand Functions

Suppose that we observe the portfolio of stocks held by institutional investor i at some time t. Let $w_{i,t,j}$ be the weight in asset $j \in \{1, ..., n_{i,t}\}$ of investor i's portfolio at time t, where $n_{i,t}$ is the number of stocks in investor i's investment universe. I follow Koijen and Yogo (2019), Koijen et al. (2023) in defining the investment universe as the set of stocks that investor i has held over the past 12 quarters. As in Koijen and Yogo (2019), Koijen et al. (2023), I normalize the portfolio weights of $w_{i,t,j}$ by the weight in the outside asset $w_{i,t,0}$. The outside asset is comprised of small stocks and stocks with missing characteristics from main specifications in Koijen and Yogo (2019), Koijen et al.

[8]For example, it is extremely challenging to identify whether a given chunk of text was written by a human or by a large language model (LLM) such as ChatGPT.

(2023).[9]

Then, the normalized demand $\delta_{i,t,j} := \frac{w_{i,t,j}}{w_{i,t,0}}$ is modelled as an exponential function of stock characteristics:

$$\delta_{i,t,j} = \frac{w_{i,t,j}}{w_{i,t,0}} = \exp\left(\gamma_{i,t} + \theta_{i,t} me_{j,t} + \beta_{i,t}^T x_{j,t}\right) \epsilon_{i,t,j} \tag{1}$$

where $\gamma_{i,t}$ is the fund-time-specific intercept, $me_{j,t}$ is the log market equity of stock j at time t, and $x_{j,t} \in \mathbb{R}^{p\times 1}$ is a vector of stock characteristics other than market equity. Coefficients $\theta_{i,t} \in \mathbb{R}^{1\times 1}$, $\beta_{i,t} \in \mathbb{R}^{p\times 1}$ correspond to the loadings of investor i's demand function at time t on log market equity me and set of stock characteristics x, respectively. Importantly, the coefficients are specific to fund i and time t. The equation (1) is estimated for each fund-quarter separately.[10]

From the econometric standpoint, the key distinction between $me_{j,t}$ and $x_{j,t}$ in Koijen and Yogo (2019) framework is that $x_{j,t}$ is assumed to be exogenous, while the endogeneity of $me_{j,t}$ is addressed via the instrumental variable approach. Specifically,

$$\mathbb{E}[me_{j,t}\,(\epsilon_{i,t,j} - 1)] \neq 0 \tag{2}$$

$$\mathbb{E}[x_{j,t}\,(\epsilon_{i,t,j} - 1)] = 0 \tag{3}$$

which states that the latent demand for asset j in (1) is correlated with the price of that asset. To address the endogeneity, I use the instrument proposed by Koijen and

[9]Specifically, the outside asset is comprised of the assets that either: 1) have CRSP share codes 12, 18; 2) are below 10th CRSP percentile in terms of market equity; 3) have missing share code, return, market equity; 3) have missing value for at least one of the characteristics used in the main specifications of Koijen and Yogo (2019), Koijen et al. (2023) (taking a superset of the sets of characteristics used in both studies, these characteristics are: book equity, market beta, foreign sales, operating profitability, sales-to-book ratio, dividend-to-book ratio, asset growth, lerner index); 4) have missing SIC 2-digit industry code.

[10]In Koijen and Yogo (2019), equation (1) is estimated quarterly, while in Koijen et al. (2023), the same equation is estimated annually (one θ_i, β_i per year) with quarter fixed effects.

Yogo (2019):[11]

$$me^{IV}_{i,t,j} = log\left(\sum_{l\neq i} AUM_{l,t}\frac{1_{l,t,j}}{1+\sum_{m=1}^{n_{i,t}} 1_{l,t,m}}\right) \tag{4}$$

where $1_{l,t,j}$ is the indicator function which is equal to 1 if asset j is in investor l's investment universe at time t, and zero otherwise. From economic standpoint, $me^{IV}_{i,t,j}$ is the counterfactual market equity of asset j if the institutional investors held all assets within their investment universe in equal weights. The identifying assumption is that the investment universe – defined as the set of stocks held by institutional investors over past 12 quarters – is determined by the investment mandate stemming from the contractual arrangements between investment vehicle and its clients, and thus, it is not related to the current latent demand of investor. For the set of investors l over which the counterfactual market equity is computed, I follow Koijen and Yogo (2019), Koijen et al. (2023) by using the entire set of 13F institutional investors.

As a benchmark estimator of investor demand function, I take the two-step ridge-IV estimator proposed in Koijen et al. (2023). In the first step, the group-level estimates $\hat{\theta}_{g,t}$, $\hat{\beta}_{g,t}$ are estimated using standard, unpenalized nonlinear GMM on the sample of grouped investors (group-level sample) under the following moment conditions:

$$\mathbb{E}\left[z_{j,t}\left(\delta_{i,t,j}\exp(-\gamma_{i,t}-\theta_{g,t}me_{j,t}-\beta^T_{g,t}x_{j,t})-1\right)\right]=0 \tag{5}$$

where $\delta_{i,t,j} := \frac{w_{i,t,j}}{w_{i,t,0}}$ and $z_{j,t} = \left(me^{IV}_{i,t,j}\ x_{j,t}\right)^T \in \mathbb{R}^{(p+1)\times 1}$. Grouping institutional investors together allows to increase the number of data points available for estimation from $n_{i,t}$ to $n_{g,t}$, thus mitigating the potential overfitting by inducing a bias towards the estimates obtained at the group level in the first step (5). Koijen et al. (2023) group investors by type (mutual funds, insurance companies) and size (AUM) so that each group has at least 2000 (including zero holdings) observations across 4 quarters.

The second-step GMM is estimated at the individual investor level with bias

[11] While alternative approaches to the IV estimation of $\theta_{i,t}$ exist (see van der Beck (2021), Huebner (2023)), I use the one proposed by Koijen and Yogo (2019) to make the results in the cross-out-of-sample exercise more comparable.

towards group-level estimates imposed by ridge penalty. The corresponding moment conditions are:

$$\mathbb{E}\left[x_{t,j}\left(\hat{\delta}_{i,t,j}\exp(-\gamma_{i,t}-\beta_{i,t}^{T}x_{t,j})-1\right)\right]-\Lambda\left(\beta_{i,t}-\hat{\beta}_{g,t}\right)=0 \tag{6}$$

Importantly, in (6), Koijen et al. (2023) impose infinite shrinkage on $\beta_{i,t}$ by forcing $\hat{\theta}_{i,t}=\hat{\theta}_{g,t}$. This infinite shrinkage is imposed by deflating $\hat{\delta}_{i,t,j}=\delta_{i,t,j}/\exp(\hat{\theta}_{g,t}me_{j,t})$, where g is the group to which investor i belongs to. Then, since $x_{j,t}$ are assumed to be exogenous, the instruments in (6) are the exogenous characteristics themselves. Another imported detail to note is that no shrinkage is applied to the intercepts $\gamma_{i,t}$ to allow investor-specific intercepts.

2.2 Data Augmentation with Synthetic Assets

By shrinking the estimates of individual demand functions of institutional investors towards the demand functions estimated on the grouped sample, Koijen et al. (2023) mitigate the issues with convergence and imprecision of estimates arising in estimation of (1).[12] However, this improvement comes at the cost of biasing the demand function estimates of individual institutional investors towards the group-level estimates.

In this paper, I seek to overcome the need of the bias towards group-level estimates in Koijen et al. (2023) by adapting a technique of data augmentation, which is commonly used in modern deep learning to prevent overfit in extremely overparameterized models. The cornerstone of the data augmentation is creation of synthetic observations based on some pre-defined, transparent, and data-driven rule.

I propose a novel, economically tractable approach to data augmentation of investor holdings data that 1) is deterministic in that it doesn't require the generation of random noise; 2) doesn't impose ex-ante economic model on the decision making process by investors; 3) tractably and effectively mitigates issues arising from overfitting and multicollinearity, even in settings where number of characteristics p is much larger than number of original holdings n.

[12] In the preceding study, Koijen and Yogo (2019) estimate only the first step (5), providing only the group-level estimates.

2.2.1 Characteristic-basis synthetic assets

To fix ideas, denote by $X_{i,t} \in \mathbb{R}^{n_{i,t} \times p}$ the matrix containing the characteristics of "original", observed in the actual data assets.[13] Here, $n_{i,t}$ corresponds to the number of assets in investor i's investment universe at time t. Further, let $\delta_{i,t} \in \mathbb{R}^{n_{i,t} \times 1}$ be the vector of the investor i's demand observed at time t.

I construct the synthetic assets as loadings on a basis of the space of asset characteristics:

$$\Psi_{i,t} := \lambda_{i,t}^{synth} \cdot \mathrm{I}_p \in \mathbb{R}^{p \times p} \tag{7}$$

where p is the dimension of the vector of asset characteristics. Throughout the paper, I refer to $\Psi_{i,t}$ as *characteristic-basis synthetic assets*. The scalar $\lambda_{i,t}^{synth}$ corresponds to the loadings of synthetic assets on the basis of asset characteristics. Characteristic-basis synthetic assets have a simple economic interpretation: k-th synthetic asset is the asset that has a loading of $\lambda_{i,t}^{synth}$ on k-th characteristic, and zero otherwise. In empirical applications in section 5, the hyperparameter $\lambda_{i,t}^{synth}$ will be selected using cross-validation, which is a common approach to selection of hyperparameters in machine learning and statistics.

Unlike synthetic data points commonly used in the computer science and statistics literature – where synthetic assets have a randomly generated component (such as random noise) – the *characteristic-basis synthetic assets* are deterministic. The latter is a useful property for applications in financial economics since the estimated via data augmentation parameters will *not* depend on the sequence of randomly drawn values. For instance, injecting randomly generated noise in the estimation of the individual demand functions of (hypothetical) fund A and fund B might lead to variation in the

[13] $X_{i,t}$ is investor i-specific due to the difference in $n_{i,t}$ across investors. Note that, however, all assets in economy $x_{t,j} \in \mathbb{R}^{p \times 1}$, $j = 1, ..., J$ are common for all investors and hence, do not have a subscript i.

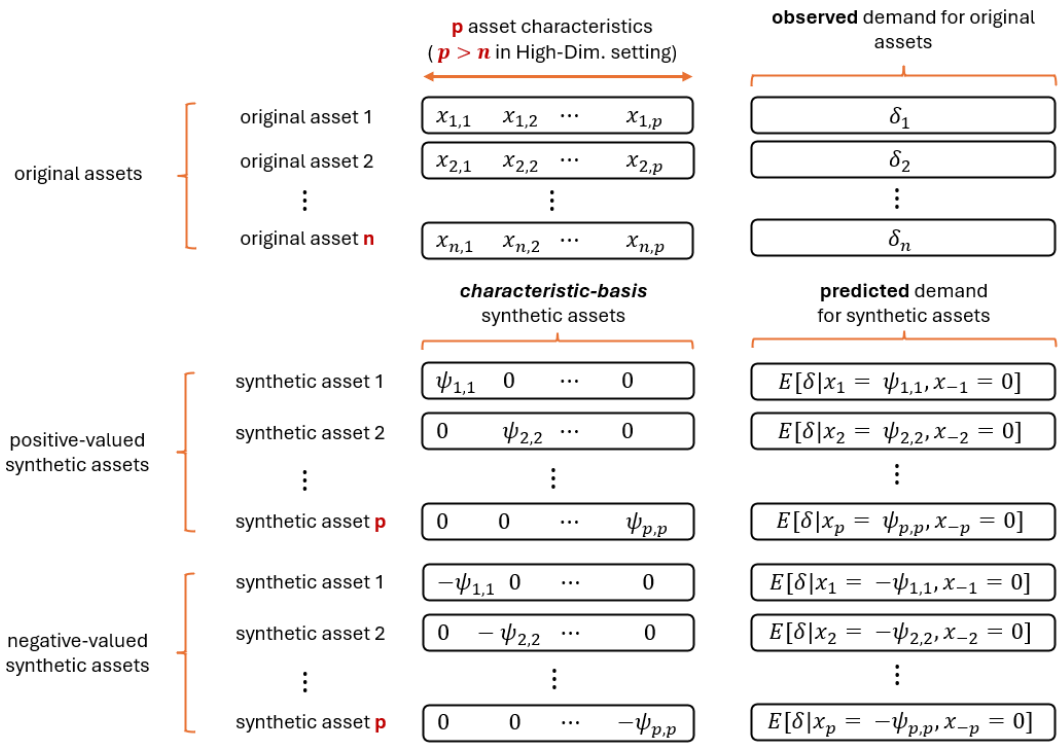


Figure 3: Data augmentation of investor's holdings

Note: Notation x_{-k} means that vector of asset characteristics x_{-k} includes all characteristics except k.

estimated demand function across researchers.[14]

Further, I construct the synthetic demand of investor i at time t for j-th synthetic asset $\psi_{i,t,j} \in \mathbb{R}^{p\times 1}$ as:

$$\delta^{synth}_{i,t,j} := \mathbb{E}[\delta_{i,t,j}|x_{t,j} = \psi_{i,t,j}] \tag{8}$$

Essentially, synthetic demand is generated as the conditional expectation of investor i's demand for synthetic assets at time t. Then, combining together (7) and (8),

[14]Common approach to ensure reproducibility for *the same* code is to set a seed for the random number generator. However, if the code among the two researchers is different, so will be the generated random numbers. To see that, suppose we have researcher 1 estimating demand functions for the sequence of funds A, B, C. If researcher sets the seed, as long as the sequence of funds A, B, C remains the same, the results well be the same. However, if the same researcher decides to estimated demand functions for A, C, B, the results will be the same for A (since the seed is the same), but different for C and B. Clearly, there is no fundamental reason why the output of the demand function estimation should depend on the sequence of funds in the *for loop*.

the *augmented dataset* is defined as:

$$X_{A,i,t} := \begin{bmatrix} X_{i,t} \\ \Psi_{i,t} \\ -\Psi_{i,t} \end{bmatrix} \in \mathbb{R}^{(n_{i,t}+2p)\times p} \ , \ \delta_{A,i,t} := \begin{bmatrix} \delta_{i,t} \\ \mathbb{E}[\delta_{i,t}|X = \Psi_{i,t}] \\ \mathbb{E}[\delta_{i,t}|X = -\Psi_{i,t}] \end{bmatrix} \in \mathbb{R}^{(n_{i,t}+2p)\times 1} \quad (9)$$

where the $p \times p$ matrix of characteristic-basis synthetic assets $\Psi_{i,t}$ is used twice: first, as the *positive-valued* synthetic assets $\Psi_{i,t}$, and second, as the *negative-valued* synthetic assets $-\Psi_{i,t}$. Intuitively, this design with dual positive-negative synthetic assets is necessary to obtain a symmetric shrinkage effect. Below, in Lemma 2, I formally show that the shrinkage effect is asymmetric if only positive-valued or only negative-valued synthetic assets are used in data augmentation. Figure 3 illustrates the construction of the augmented dataset in (9).

2.2.2 Prediction of Synthetic Demand for Synthetic Assets

From 9, in order to construct investor i's synthetic demand at time t for synthetic assets $\Psi_{i,t}$, one needs to specify the conditional expectation function $\mathbb{E}[\delta_{i,t,j}|x_{t,j} = \psi_{i,t,j}]$. Given the specification of demand function in (1), one natural choice for this conditional expectation function is:

$$\hat{\delta}_{i,t,j}^{synth} = \exp\left(\left(\hat{\beta}_{i,t}^{target}\right)^T \psi_{i,t,j}\right) \quad (10)$$

where $\hat{\beta}_{i,t}^{target}$ is some prior about investor i's time-t demand function coefficients, estimated from the original dataset $(X_{i,t}, \delta_{i,t})$. The issue with this specification of synthetic demand is that one has to provide a good-quality prior estimate $\hat{\beta}_{i,t}^{target}$, and estimating (10) is challenging due to the exact same problems of overfit and multicollinearity. Therefore, one has to reduce the complexity of (1) to obtain $\hat{\beta}_{i,t}^{target}$.

To make construction of (10) empirically feasible, I propose to reduce complexity of (1) along two dimensions. First, instead of fitting a nonlinear GMM with iterative

algorithm, I estimate $\beta_{i,t}^{target}$ via a more robust log-linear specification[15] of (1):

$$\log\left(\delta_{i,t,j}\right) = \gamma_{i,t} + \left(\beta_{i,t}^{target}\right)^T x_{j,t} + \eta_{i,t,j} \tag{11}$$

Second, I impose a ridge penalty on the deviation of $\beta_{i,t}^{target}$ from $p \times 1$ vector of zeros. In this way, the values of $\beta_{i,t}^{target}$ are shrunk towards zero, reducing the effective complexity of the model.[16] Then, the $\beta_{i,t}^{target}$ are estimated as:

$$\hat{\beta}_{i,t}^{target} = \left(X_{i,t}^T X_{i,t} + \lambda_{i,t}^{simple} \cdot I_p\right)^{-1} \left(X_{i,t}^T \log(\delta_{i,t})\right) \tag{12}$$

where $\lambda_{i,t}^{simple}$ is the hyperparameter governing the strength of shrinkage of $\beta_{i,t}^{target}$ to zero. The intercept $\gamma_{i,t}$ is fitted by de-meaning $\log(\delta_{i,t})$ and $X_{i,t}$. As a special case, when $\lambda_{i,t}^{simple} = 0$, the simplified ridge becomes OLS:

$$\hat{\beta}_{i,t}^{target} = \left(X_{i,t}^T X_{i,t}\right)^{-1} \left(X_{i,t}^T \log(\delta_{i,t})\right) \text{ for } \lambda_{i,t}^{simple} = 0 \tag{13}$$

To ensure that $\lambda_{i,t}^{simple}$ reflects the data generating process of investor i's demand at time t, the $\lambda_{i,t}^{simple}$ is fitted adaptively via cross-validation for each investor i-time t pair separately.

As can be noticed from (9), the framework of holdings augmentation with synthetic assets is more general than a specific choice of the synthetic demand for synthetic assets chosen in (12). Specifically, in the paper below I refer to data augmentation with log-linearized ridge targets in (12) as augmenting with *solo* synthetic assets. I also denote by *dual* synthetic assets augmentation two sets of synthetic assets for which the first set of synthetic assets is assigned synthetic demand according to the log-linearized ridge targets, while the second set of synthetic assets is assigned the synthetic demand of zero. The total number of synthetic assets then becomes $2p + 2p = 4p$, and the relative strength of the penalization with synthetic assets is then chosen via cross-validation. While augmenting with solo synthetic assets allows to shrink the final GMM model estimates towards linear ridge targets, the second set of zero-demand synthetic assets implements shrinkage of coefficients towards zero, further improving the robustness of

[15] $\log(\delta_{i,t,j})$ is a natural logarithm of $\delta_{i,t,j} = \frac{w_{i,t,j}}{w_{i,t,0}}$ over the subset of strictly positive portfolio weights $w_{i,t,j} > 0$.

[16] See Hoerl and Kennard (1970) or van Wieringen (2023) for textbook treatment.

the model towards overfitting. As will be shown in section 5, the models with dual synthetic assets provide the highest cross-out-of-sample performance.

2.2.3 Properties of Data Augmentation with Synthetic Assets

Having defined the data augmentation with characteristic-basis synthetic assets, I turn to the formal description of effect of such data augmentation on the GMM estimation of (1). This section is technical and is not essential for understanding of the key intuition behind data augmentation. Those readers who are interested in application of the method can skip this subsection.

For ease of exposition in this subsection, assume all regressors are exogenous, and that true intercept in DGP of $\hat{\delta}_{i,t,j}$ is normalized to 1. The extensions to IV estimation are described in the next subsection. The moment condition corresponding to the nonlinear GMM estimation of (1) are:

$$\mathbb{E}\left[z_{i,t,j}\left(\hat{\delta}_{i,t,j}\exp(-x_{i,t,j}^{T}\beta_{i,t})-1\right)\right]=0 \tag{14}$$

The sample counterpart of (14) on original dataset:

$$\frac{1}{n}\sum_{j=1}^{n}\left[z_{i,t,j}\left(\hat{\delta}_{i,t,j}\exp(-x_{i,t,j}^{T}\beta_{i,t})-1\right)\right]=0 \tag{15}$$

Proposition 1 describes estimation of nonlinear GMM under moment conditions (14) on the augmented dataset with synthetic assets (9) and synthetic demand for synthetic assets (10).

Proposition 1. *Augmentation of a nonlinear demand function estimator in (1) under moment conditions (14) with characteristic-basis synthetic assets defined as in (9), (10) is equivalent to the nonlinear GMM with nonlinear penalty on the deviation of coefficients $\beta_{i,t}$ from the target $\beta_{i,t}^{target}$:*

$$\hat{\mathbb{E}}\left[z_{A,i,t,j}\left(\hat{\delta}_{A,i,t,j}\exp(-x_{A,i,t,j}^{T}\beta_{i,t})-1\right)\right]=0$$

$$\Leftrightarrow$$

$$\frac{1}{n}\sum_{j=1}^{n}\left[z_{i,t,j}\left(\hat{\delta}_{i,t,j}\exp(-x_{i,t,j}^{T}\beta_{i,t})-1\right)\right]+\pi_{synth}\left(\lambda_{i,t}^{synth},\ \beta_{i,t}-\beta_{i,t}^{target}\right)=0$$

with:

$$\pi_{synth}\left(\lambda_{i,t}^{synth},\ \beta_{i,t}-\beta_{i,t}^{target}\right)=\frac{1}{n}\cdot\lambda_{i,t}^{synth}\left(\exp(\lambda_{i,t}^{synth}(\beta_{i,t}^{target}-\beta_{i,t}))-\exp(\lambda_{i,t}^{synth}(\beta_{i,t}-\beta_{i,t}^{target}))\right)$$

where exp() is applied element-wise, and $\lambda_{i,t}^{synth}\geq 0$ *is a hyper-parameter governing the strength of the shrinkage via synthetic assets, and the k-th moment condition-specific penalty* $\pi_{k}^{synth}\in\mathbb{R}^{1\times 1}$ *obeys the following:*

$$\begin{aligned}
\lambda_{i,t}^{synth}=0\ \Rightarrow\forall k\in\{1,...,p\},\ \pi_{k}^{synth}\left(\lambda_{i,t}^{synth},\ \beta_{k,i,t}-\beta_{k,i,t}^{target}\right)=0\\
\beta_{k,i,t}=\beta_{k,i,t}^{target}\ \Rightarrow\ \pi_{k}^{synth}\left(\lambda_{i,t}^{synth},\ \beta_{k,i,t}-\beta_{k,i,t}^{target}\right)=0\\
\forall\beta_{k,i,t}\neq\beta_{k,i,t}^{target},\ as\ \lambda_{i,t}^{synth}\to+\infty\ ,\ \left\|\pi_{k}^{synth}\left(\lambda_{i,t}^{synth},\ \beta_{k,i,t}-\beta_{k,i,t}^{target}\right)\right\|_{2}\to+\infty\\
\forall\lambda_{i,t}^{synth}>0,\ as\ |\beta_{k,i,t}-\beta_{k,i,t}^{target}|\to+\infty\ ,\ \left\|\pi_{k}^{synth}\left(\lambda_{i,t}^{synth},\ \beta_{k,i,t}-\beta_{k,i,t}^{target}\right)\right\|_{2}\to+\infty
\end{aligned}$$

Proof: see Appendix C.3.

There are two key takeaways from Proposition (1). First, GMM estimation on the dataset augmented with synthetic assets is equivalent to the GMM estimation on the original dataset with a cost imposed on the deviation of parameters $\beta_{i,t}$ from some pre-specified by empiricist target. This way, GMM with data augmentation effectively acts as a shrinkage estimator, preventing overfitting by shrinking the coefficient to a pre-specified prior. Second, the GMM estimation on original dataset is a subcase of GMM estimation with data augmentation when the loadings of synthetic assets on characteristic basis goes to zero ($\lambda_{i,t}^{synth}\to 0$). Since $\lambda_{i,t}^{synth}$ is selected via cross-validation, this means that the solution to the GMM with data augmentation can be arbitrarily close to the "standard" GMM, provided that the data generating process for investor i at time t supports selection of $\lambda_{i,t}^{synth}=0$ in cross-validation.

The following two lemmas are useful to prove Proposition 1 and provide intuition on importance of using both positive- and negative-valued synthetic assets.

Lemma 1. *The penalties induced by characteristic-basis synthetic assets have the following functional form:*

a) for positive-valued synthetic assets:

$$\pi_{+}^{synth}\left(\lambda_{i,t}^{synth},\, \beta_{i,t}-\beta_{i,t}^{target}\right) = \frac{1}{n}\cdot\lambda_{i,t}^{synth}\left(\exp(\lambda_{i,t}^{synth}(\beta_{i,t}^{target}-\beta_{i,t}))-1_p\right)$$

b) for negative-valued synthetic assets:

$$\pi_{-}^{synth}\left(\lambda_{i,t}^{synth},\, \beta_{i,t}-\beta_{i,t}^{target}\right) = -\frac{1}{n}\cdot\lambda_{i,t}^{synth}\left(\exp(\lambda_{i,t}^{synth}(\beta_{i,t}-\beta_{i,t}^{target}))-1_p\right)$$

c) with all synthetics assets:

$$\pi_{synth}\left(\lambda_{i,t}^{synth},\, \beta_{i,t}-\beta_{i,t}^{target}\right) = \frac{1}{n}\cdot\lambda_{i,t}^{synth}\left(\exp(\lambda_{i,t}^{synth}(\beta_{i,t}^{target}-\beta_{i,t}))-\exp(\lambda_{i,t}^{synth}(\beta_{i,t}-\beta_{i,t}^{target}))\right)$$

where 1_p *is* $p\times 1$ *vector of ones, and* $\exp()$ *is applied element-wise.*

Proof: see Appendix C.1.

In essence, Lemma 1 derives the shape of the penalty imposed individually by positive-valued and negative-valued synthetic assets. Another important result from Lemma 1 is that the shape of the penalties induced by synthetic assets *adapt* to the functional specification of the moment condition of nonlinear GMM. Indeed, by comparing expressions for penalties provided in the lemma above to (14), one can note that exponential functional form of the GMM's moment condition resulted in the exponential form of the penalty on the coefficients. Figure 4 illustrates the shape of each of the two "branches" of penalties and plots the shape of the total penalty from both types of synthetic assets.

It is interesting to compare the shape of the total penalty induced by all synthetic assets to the shape of ridge penalty (see Figure 4). Since the penalty imposed by synthetic assets takes exponential form, the penalty grows much slower for the values of $\beta_{i,t}$ close to the target $\beta_{i,t}^{target}$ compared to the values of $\beta_{i,t}$ relatively far from the target. This adaptive shape of the penalty reflects the costs of overfitting for the exponential

functional form of GMM in (14). If $\beta_{i,t}$ in (14) is overfitted by 0.1 close to the target, the fitted values $\hat{\delta}_{i,t}$ will be less implausible than if the $\beta_{i,t}$ is overfitted by 0.1 far from the target. At the same time, the ridge penalty on the moment condition grows at the constant speed. As will is shown the in cross-out-of-sample validation exercise, the adaptive shrinkage via data augmentation outperforms standard ridge penalty.

The following lemma formally shows asymmetry of the penalties induced by positive-valued and negative-valued synthetic assets.

Lemma 2. *The penalties induced by using only positive-valued or only negative-valued characteristic-basis synthetic assets are asymmetric around the coefficient target* $\beta_{k,i,t}^{target}$, $k \in \{1, ..., p\}$. *Specifically,* $\forall \lambda_{i,t}^{synth} > 0$, *for positive-valued synthetic assets:*

$$\lim_{(\beta_{k,i,t} - \beta_{k,i,t}^{target}) \to +\infty} \left\| \pi_{+}^{synth} \left(\lambda_{i,t}^{synth},\, \beta_{i,t} - \beta_{i,t}^{target} \right) \right\|_2 = \frac{1}{n} \cdot \lambda_{i,t}^{synth}$$

$$\lim_{(\beta_{k,i,t} - \beta_{k,i,t}^{target}) \to -\infty} \left\| \pi_{+}^{synth} \left(\lambda_{i,t}^{synth},\, \beta_{i,t} - \beta_{i,t}^{target} \right) \right\|_2 = +\infty$$

For negative-valued synthetic assets:

$$\lim_{(\beta_{k,i,t} - \beta_{k,i,t}^{target}) \to +\infty} \left\| \pi_{-}^{synth} \left(\lambda_{i,t}^{synth},\, \beta_{i,t} - \beta_{i,t}^{target} \right) \right\|_2 = +\infty$$

$$\lim_{(\beta_{k,i,t} - \beta_{k,i,t}^{target}) \to -\infty} \left\| \pi_{-}^{synth} \left(\lambda_{i,t}^{synth},\, \beta_{i,t} - \beta_{i,t}^{target} \right) \right\|_2 = \frac{1}{n} \cdot \lambda_{i,t}^{synth}$$

where $\pi_{+}()$ *denotes penalty induced by positive-valued synthetic assets, and* $\pi_{-}()$ *denotes penalty corresponding to the negative-valued synthetic assets.*

At the same time, with both positive- and negative-valued synthetic assets:

$$\lim_{(\beta_{k,i,t} - \beta_{k,i,t}^{target}) \to +\infty} \left\| \pi_{synth} \left(\lambda_{i,t}^{synth},\, \beta_{i,t} - \beta_{i,t}^{target} \right) \right\|_2 = +\infty$$

$$\lim_{(\beta_{k,i,t}^{target} - \beta_{k,i,t}) \to +\infty} \left\| \pi_{synth} \left(\lambda_{i,t}^{synth},\, \beta_{i,t} - \beta_{i,t}^{target} \right) \right\|_2 = +\infty$$

Proof: see Appendix C.2.

Lemma 2 shows the reason behind the usage of both positive- and negative-valued synthetic assets. If only positive-valued synthetic assets $\Psi_{i,t}$ are used for data

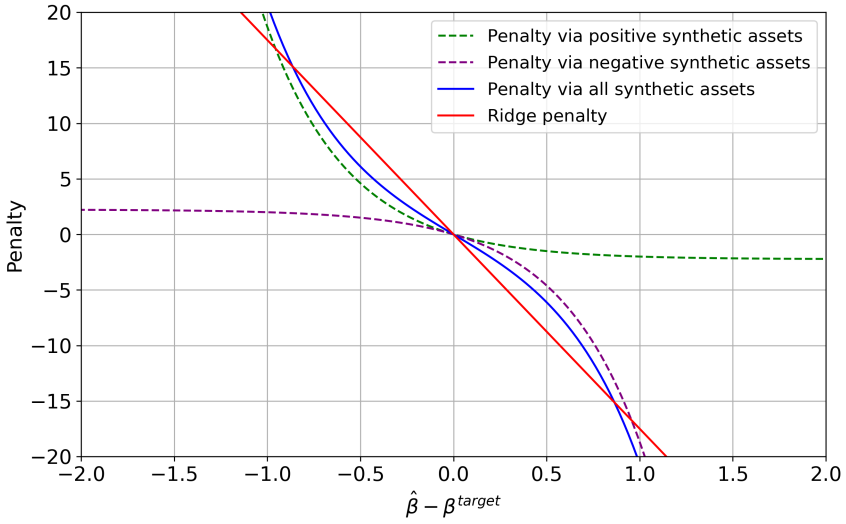


Figure 4: Data augmentation as penalty on the moment condition

Note: Quadratic ridge penalty in GMM objective corresponds to the linear penalty on the GMM's moment conditions.

augmentation, one obtains a penalty which will apply a large cost to the negative deviation of the coefficient from the target but only moderate cost on the positive deviation, even as the deviation grows very large. Likewise, if only negative-valued synthetic assets $-\Psi_{i,t}$ are used, positive deviations from the target will be associated with large cost, while negative deviations will remain relatively unpenalized. Figure 4 illustrates the intuition behind Lemma 2.

To see why this asymmetry is undesirable, consider augmentation with positive-valued synthetic assets $\Psi_{i,t}$ only. To further simply the intuition, suppose $\beta_{k,i,t}^{target} = 0$. Then, for a moderate value of $\lambda_{i,t}^{synth} << \infty$, even extremely large positive deviations of $\beta_{k,i,t}$ are permitted by the estimator. This, in turn, permits the estimator to overfit the data by choosing large positive coefficients. The only way to prevent the overfit via the large positive coefficients in this case is to set a very large $\lambda_{i,t}^{synth}$. Now, suppose that we use both positive-valued synthetic assets $\Psi_{i,t}$ and negative-valued synthetic assets $\Psi_{i,t}$. Then, for a given moderate $\lambda_{i,t}^{synth}$, both large negative and large positive coefficients will be penalized similarly.

2.3 Data Augmentation in the Presence of Endogenous Regressors

Since the core objective of data augmentation in this paper is to facilitate the estimation of complex demand functions in data-scarce settings, the two key challenges arise in IV estimation with data augmentation. The first one lies in consistent estimation of the coefficient on endogenous variable *me* in the presence of large-dimensional vector of confounding variables $x \in \mathbb{R}^{p\times 1}$. The second hurdle is the seamless integration of data augmentation and consistent estimation of coefficient on endogenous variable: Since augmentation with synthetic assets introduces shrinkage-like bias in the estimates, it is important to design the estimation algorithm in such a way that the mentioned bias does not spill over to the IV estimation. In what follows, I propose and describe the solutions to these two challenges.

To fix ideas, consider moment condition:

$$\mathbb{E}\left[z_{i,t,j}\left(\delta_{i,t,j}\exp(-me_{t,j}\theta_{g,t}-\sum_{k=1}^{p}x_{k,t,j}\beta_{k,g,t})-1\right)\right]=0 \tag{16}$$

Since the number of confounding variables p can be large compared to the number of observations n used in the estimation, the GMM estimation of (16) will typically fail due to the multicollinearity and overfit. The estimation of (16) is further aggravated by the low precision in the first stage.

In KRY23, authors propose estimation of (16) with small number of asset characteristics ($p < 10$) on the *grouped* dataset, where individual institutional investors are pooled into groups with at least 2,000 holdings (including zero holdings). Through grouping, KRY23 increase the number of available for estimation observations, while also limiting the complexity of the asset demand function to be estimated.

To estimate $\theta_{g,t}$ in the settings where the number of confounding variables is relatively large compared to the number of available observations, I employ a debiased estimation under moment conditions that are immunized to impact of large dimensionality of covariates. Specifically, I estimate the coefficient on market equity $\theta_{g,t}$ under

the following "immunized" moment condition:

$$\hat{\mathbb{E}}\left[\check{z}_{i,t,j}\left(\log(\check{\delta}_{i,t,j}) - \check{me}_{t,j}\theta_{g,t}\right)\right] = 0 \tag{17}$$

which yields the following estimator:

$$\hat{\theta}_{g,t} = \left(\hat{\mathbb{E}}\left[\check{z}_{i,t,j}\check{me}_{t,j}\right]\right)^{-1}\hat{\mathbb{E}}\left[\check{z}_{i,t,j}\log(\check{\delta}_{i,t,j})\right] \tag{18}$$

where:

$$\check{z}_{i,t,j} = z_{i,t,j} - \check{\mathbb{E}}\left[z_{i,t,j}|x_{t,j}\right] \tag{19}$$

$$\check{me}_{t,j} = me_{t,j} - \check{\mathbb{E}}\left[me_{t,j}|x_{t,j}\right] \tag{20}$$

$$\check{\delta}_{i,t,j} = \delta_{i,t,j} - \check{\mathbb{E}}\left[\delta_{i,t,j}|x_{t,j}\right] \tag{21}$$

To allow conditional expectation functions $\mathbb{E}[v_{i,t,j}|x_{t,j}]$ to be arbitrarily dense in $x_{t,j} \in \mathbb{R}^{p\times 1}$, I estimate it via Partial Least Squares (PLS) using algorithm of de Jong (1993).

3 Simulation Study

To verify the methodology developed in the previous section, I test its performance on simulated investor holdings data. Specifically, I simulate $\mathcal{N}_{sim} = 256$ mutual funds whose holdings follow the data generating process (hereafter, DGP) implied by the demand function specification (1) from Koijen and Yogo (2019), Koijen et al. (2023). Namely, the market equity me is correlated with latent demand ϵ, while all other asset characteristics $x \in \mathbb{R}^{p\times 1}$ are exogenous. Each fund has $n = 250$ assets in its investment universe. The dimension of exogenous characteristics p is varied across 6 versions of the simulated institutional investor universes: $p \in \{8, 25, 50, 100, 150, 250\}$. The case $p = 8$ corresponds to the low-dimensional case where p is relatively small compared to the number of assets in funds' investment universe $n = 250$. Versions of simulation with large p correspond to the case with high-dimensional demand functions. Due to the issues with multicollinearity and overfitting arising in high-dimensional estimation, the higher p, the more challenging is the estimation. The detailed description of the

simulation design in provided in Appendix B.

Having simulated the holdings, I compare the relative performance of the demand function estimators based on data augmentation versus the two-step ridge-GMM estimator with shrinkage towards group priors proposed in Koijen et al. (2023). The simulation shows that in the presence of heterogeneity across investors, estimation of demand functions using GMM with data augmentation leads to more precise estimates of the demand function loadings $\beta \in \mathbb{R}^{p\times 1}$. I measure the precision of estimation as a mean squared error (MSE) of estimated $\hat{\beta}$ compared to the true β^{true}. Since in simulated data, one knows β^{true}, computation of this MSE is trivial. The results of simulation are shown in Figure 5.

3.1 Design of Simulation Study

In this subsection, I describe in detail the design of simulation study. Readers who are not interested in technical details can skip this subsection and continue with section 3.2.

3.2 Results of Simulation Study

To evaluate the performance of competing estimators, I compute the mean squared error of estimated demand function loadings. Given $\hat{\beta}^M$ obtained from estimator M, the mean squared error for a given dimension of the demand function p is computed over the entire simulated universe of $\mathcal{N}_{sim}$ investors:

$$MSE(\hat{\beta}^M, p) = \frac{1}{\mathcal{N}_{sim}} \sum_{i=1}^{\mathcal{N}_{sim}} \frac{1}{p} \sum_{k=1}^{p} (\hat{\beta}_k^M - \beta_k^{true})^2 \tag{22}$$

As alternative metric of performance, I also compute mean absolute error (mean absolute deviation):

$$MAE(\hat{\beta}^M, p) = \frac{1}{\mathcal{N}_{sim}} \sum_{i=1}^{\mathcal{N}_{sim}} \frac{1}{p} \sum_{k=1}^{p} |\hat{\beta}_k^M - \beta_k^{true}| \tag{23}$$

Figure 5 presents the results of simulation study. Estimation of demand func-

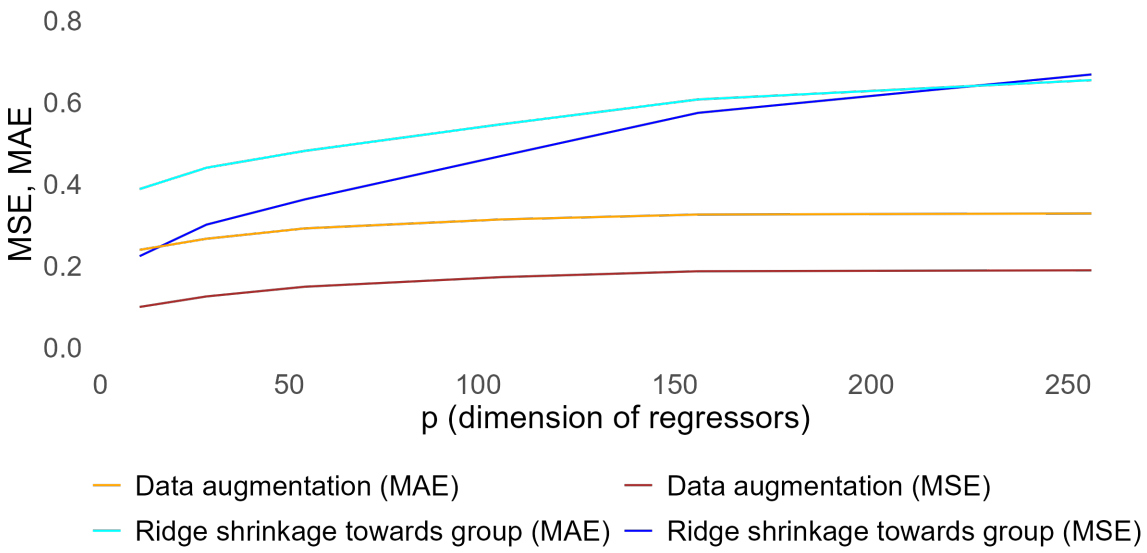


Figure 5: Results of simulation study

Note: "Ridge shrinkage towards group" corresponds to the estimator proposed by Koijen et al. (2023). "Data augmentation" is the nonlinear GMM with synthetic assets with dual synthetic assets proposed in section 2 but without shrinkage towards group targets.

tions with data augmentation consistently outperforms the ridge-GMM estimation with shrinkage towards group targets in terms of the estimation precision of $\hat{\beta}$ according to both mean squared error and mean absolute error. The intuition behind the finding is that in the presence of heterogeneity across investors, estimating individual fund-level demand functions with shrinkage towards group targets is likely to induce a substantive bias into the estimates, leading to poor MSE despite lower variance. On the other hand, by generating synthetic assets from a simplified demand function, one can achieve effective regularization and mitigate overfitting without inducing the bias towards the group. Notably, the simulation relies on the assumption that in the true data generating process (DGP), there is significant heterogeneity across investors. In section 5, I will show on the real mutual funds data that, indeed, relying on shrinkage towards group-level estimates leads to substantial deterioration of the estimation quality, even when compared to na"ive targets where priors on all demand function loadings is assumed to be zero.

4 Data

4.1 Holdings of Institutional Investors

In this study, I estimate demand functions for *active* mutual funds only.[17] There are three main reasons for this. First, for mutual funds, we observe holdings at the individual fund level, rather than at the management company level as in 13F data. This provides a natural laboratory to study differentiation as we not only observe what management companies (such as Fidelity or AQR) do, but also what are the portfolio choices of the individual funds within the fund families. Second, active mutual funds is a large group of institutional investors who have strong incentives to deliver well-performing strategies since mutual fund flows tend to follow funds' past performance. Third, for active mutual funds, I am able to exploit the Morningstar mutual fund ratings reform in 2002 to instrument for the competition-driven increase in flows to mutual funds.

The sample used in the cross-out-of-sample analysis spans from January 1990 to December 2022. The sample used for the empirical analysis of the effect of increased competition on mutual fund differentiation start on January 2001 and ends on December 2007. The start of the sample is set to avoid the dot-com bubble, while the end of the sample is set before the Global Financial Crisis. Since the Morningstar mutual fund ratings reform was implemented on June 2002, the limits of the sample provide sufficient pre-shock period to check for parallel trends, as well as more than 4 years of data after the reform to study long-term effects of the changes.

Following the recommendations in Zhu (2020), I use the mutual fund holdings data from Thomson Reuters s12 for the first part of the sample, and CRSP Mutual Fund database for second part of the sample. The switch date for the two datasets is set to January 2010, which is roughly the time when CRSP mutual fund holdings data becomes high-quality. This switch allows to avoid issues with missing new funds in Thomson Reuters dataset described in Zhu (2020). I construct an overarching mutual fund identifier $portf\_id$ based on CRSP's $crsp\_cl\_grp$ and Thomson Reuters' $wficn$. To seamlessly switch the holdings source for funds that exist in both CRSP and Thomson

[17] However, the instrument for market equity is constructed using the counterfactual portfolio weights of entire 13F universe of institutional investors

Reuters, I match $crsp\_cl\_grp$ to $wficn$ based on the *MF Links* map between $wficn$ and $crsp\_fundno$, which is later matched with $crsp\_cl\_grp$ through the $crsp\_cl\_grp$-$crsp\_fundno$ map provided in CRSP Mutual Fund database's fund summary file. Mutual fund holdings from the CRSP are merged to the overarching fund identifier $portf\_id$ through $crsp\_cl\_grp$-$crsp\_portno$ mapping, while the holdings from Thomson Reuters s12 file are merged via the mapping between Thomson Reuter's $fundno$ and $wficn$ from *MF Links*.

I provide the descriptive statistics on the number of stocks held by active equity mutual funds in Table A.1. For comparison, I tabulate the same statistics for passive equity mutual funds in Table A.3 and 13F institutional investors in Table A.2.

4.2 Stock Characteristics

To determine the list of stock characteristics to be used in the high-dimensional demand function specification, I start with the list of stock characteristics that comprise "factor zoo" in Jensen et al. (2023). Then, I remove characteristics that are likely to be endogenous in the demand function specification (1). For example, I remove return- and price-based signals such as those comprising "momentum", "seasonality", and "short-term reversal" themes. Some stock characteristics in "factor zoo" are constructed by scaling a given variable by market equity me. In such cases, I re-scale the characteristic by be to avoid endogenous variation in me that could violate the exclusion restriction. This way, only me in (1) is endogenous and requires an instrument. The list of stock characteristics (other than me and industry dummies) used in this study is provided in Table A.4. The correlation matrix between stock characteristics is provided in Figure A.2. In Figure A.3, I plot the correlation matrix between stock characteristics and 85 SIC 2-digit dummies.

In settings with high-dimensional estimation, it is customary to rank-normalize the regressors.[18] To preserve the original interpretation of the coefficient β_0 that describes price elasticity in Koijen and Yogo (2019), Koijen et al. (2023), I do not rank-normalize market equity but include it in logs $me := log(ME)$, as in Koijen and Yogo

[18]The reason for this rank-normalization (or any alternative standardization) comes from the fact that regularized estimators are typically not invariant to the scale of the inputs. Once the scale of the inputs has been normalized across all regressors, this issue is avoided.

(2019), Koijen et al. (2023). All other stock characteristics in are rank-normalized to the interval of $[0, 1]$ except for SIC 2-digit dummies which are either 0 or 1 by construction. To ensure that this rank-normalization does not affect price elasticity estimates other than thought the covariance with characteristics, I estimate demand functions with log *me* and rank-normalized log *be* rather than log market-to-book.[19]

5 Empirical Performance of Demand Functions Estimators

5.1 Cross-out-of-sample validation

In this section, I compare the performance of competing estimators of investor demand functions on equity holdings data of active mutual funds. First, I estimate the mutual fund demand functions using ridge-GMM with shrinkage towards group-level estimates developed Koijen et al. (2023). Then, I estimate the demand functions using a simple alternative that shrinks the estimates towards the $p \times 1$ vector of zeros, where p is the dimension of characteristic vector. This estimator is a nonlinear counterpart of the standard ridge regression where shrinkage the coefficients towards zero is applied to reduce the variance of the estimates, mitigate multicollinearity, and ensure that estimates are well-defined. Unlike the original estimator in Koijen et al. (2023), this estimator does *not* rely on shrinkage towards the group-level estimates and therefore, it does not mask the heterogeneity of the demand functions' estimates across institutional investors. However, by shrinking coefficients towards zero, it understates the magnitude of characteristic-specific demand elasticities. Finally, I estimate mutual fund demand functions using the proposed in the section 2 method of estimation of demand system via data augmentation. Since the method in Koijen and Yogo (2019) allows one to estimate only group-level demand loadings, I use Koijen et al. (2023) as the main benchmark. Essentially, the key difference between Koijen and Yogo (2019) and Koijen

[19]Note that for $me := \log(ME)$, $be := \log(BE)$, and $mb := \log(ME/BE)$:

$$\begin{aligned}\beta_{0,i,t} me_{j,t} + \beta_{1,i,t} be_{j,t} =& \beta_{0,i,t} \left(me_{j,t} - be_{j,t}\right) + \left(\beta_{0,i,t} + \beta_{1,i,t}\right) be_{j,t} \\ =& \beta_{0,i,t} mb_{j,t} + \left(\beta_{0,i,t} + \beta_{1,i,t}\right) be_{j,t}\end{aligned}$$

Hence, we can estimate $\beta_{0,i,t}$ either as a coefficient on *me* or *mb* – the only thing that is affected by this choice is the coefficient on *be*.

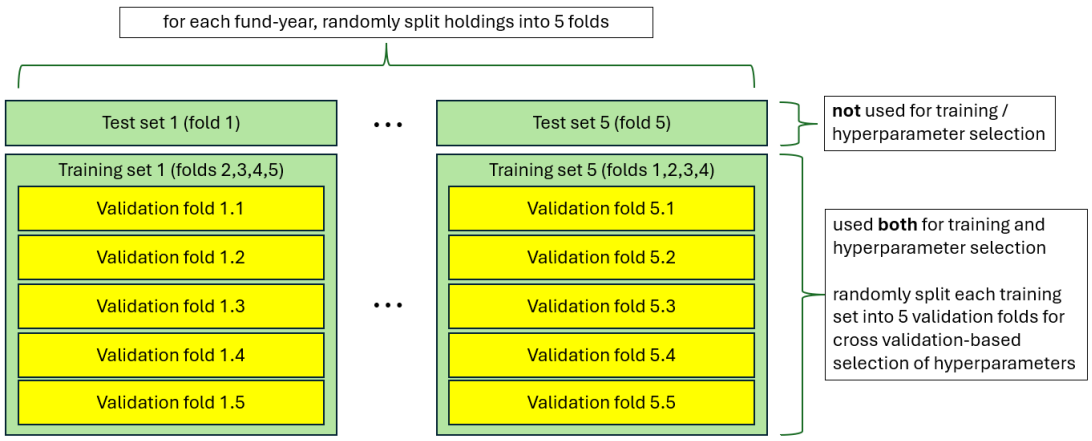


Figure 6: Illustration of the Cross-Out-of-Sample Analysis

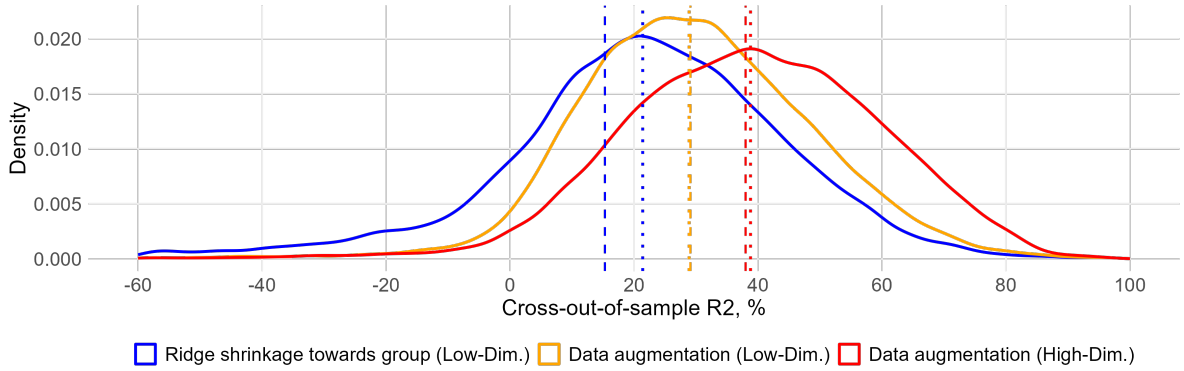
Training set includes folds other than the test fold, and cross-validation is performed only on the training set to ensure that the prediction on the test fold is a proper out-of-sample exercise. The data is split on the 5 folds randomly *within* each fund-year set of fund holdings.

et al. (2023) is that the latter method performs a second-step estimation where the group-level estimates are used as shrinkage targets.

For all estimators (including those based on data augmentation), the price elasticity is estimated on the grouped dataset to ensure the sufficient strength of the first stage in IV estimation (while the coefficients on all other stock characteristics are estimated using only individual fund-specific holdings data). The two-step nonlinear ridge-GMM with group targets uses the IV estimates obtained nonlinear GMM in (5), while the nonlinear ridge-GMM with zero targets and GMM with data augmentation use estimates of price elasticity obtained from linear debiased GMM described in section 2.3. Overall, the estimates of price elasticities are highly correlated across three methods. The AUM-weighted average of the mutual fund price elasticities is about 0.5, which is similar to the estimates obtained for small active 13F investors by Koijen et al. (2023). Since price elasticities are not crucial for my application, I provide the details for the interested readers in Appendix D.

I compare the performance of competing demand function estimators through the cross-out-of-sample analysis. First, for each fund-quarter (e.g. 2022-Q2 of the Fidelity Magellan Fund), the data is randomly split into 5 folds. Then, for each fold $k = 1, ..., 5$, I perform estimation and hyperparameter selection using only the data from the 4 folds other than k, and then use the data from the fold k to construct the out-of-sample predictions of the fund portfolio weights $\hat{w}$ and compute the out-of-sample error $w - \hat{w}$.

Panel A: Distribution of cross-out-of-sample R^2 across fund-quarters



Panel B: Time-series variation in median cross-out-of-sample R^2

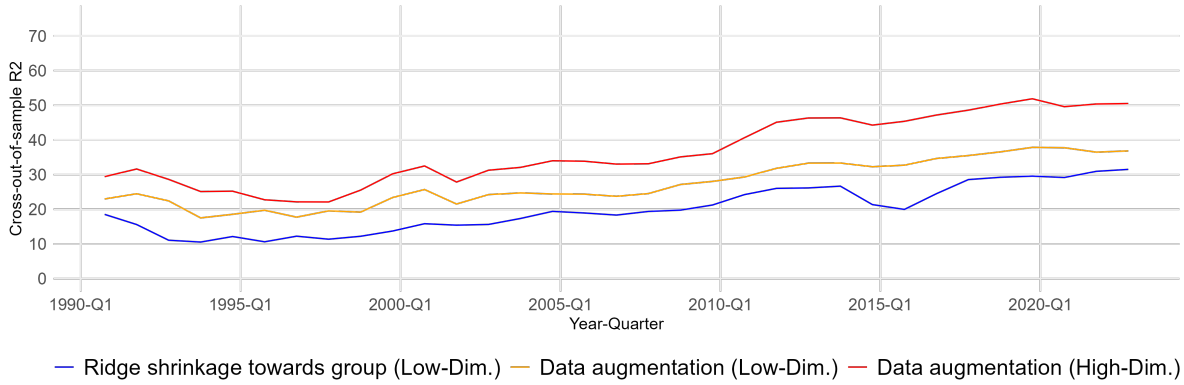


Figure 7: Performance of demand function estimators

Note: Panel A shows the distribution of cross-out-of-sample R^2 across fund-quarters. Vertical dashed (dotted) lines show the mean (median) of the distribution of cross-out-of-sample R^2 for each estimation approach. Panel B plots the time-series of median cross-out-of-sample R^2. The figure shows only the low-dimensional version of the nonlinear GMM with shrinkage towards group target since for high-dimensional demand functions, the estimator systematically does not converge. In line with Koijen et al. (2023), the demand functions are estimated for each fund annually. To mitigate the impact of outliers on the mean, I winsorize the distribution of R^2 for each estimator at 2.5%.

When performing cross-validation, I randomly split each training set into 5 validation folds on which the CV-based search of hyperparameters is performed. This nested structure (cross-validation within training set) ensures that prediction on the test set is a proper out-of-sample exercise. Once the out-of-sample errors are obtained for all 5 folds, I compute the OOS MSE (mean squared error) and OOS R^2 for each method using the errors from all 5 folds:

$$MSE_{OOS,i,t}(M) = \frac{1}{N_{i,t}} \sum_{j=1}^{N_{i,t}} \left(w_{i,j,t}^{test} - \hat{w}_{i,j,t}^{pred} \right)^2$$

Table 1: Distribution of cross-out-of-sample R^2 across fund-years

Estimator	mean	sd	p10	p25	p50	p75	p90	Diff(mean)	t-stat
Nonlinear ridge-GMM with group target (Low-dim.)	15.31	36.20	-16.02	7.00	21.40	35.51	48.42		
Nonlinear ridge-GMM with zero target (Low-dim.)	26.47	19.95	5.04	15.15	26.66	39.29	51.61	11.15	51.25
Nonlinear ridge-GMM with linear target (Low-dim.)	25.80	21.57	4.04	14.77	26.51	39.37	51.71	10.50	46.57
Nonlinear ridge-GMM with dual target (Low-dim.)	28.42	18.57	6.51	16.29	28.06	40.84	53.12	13.12	57.25
Nonlinear GMM with single synthetic assets (Low-dim.)	28.00	19.80	5.53	16.17	28.23	41.15	53.28	12.69	56.94
Nonlinear GMM with dual synthetic assets (Low-dim.)	29.10	18.68	6.87	16.85	28.89	41.73	53.86	13.79	60.69
Nonlinear ridge-GMM with zero target (High-dim.)	28.07	20.78	6.52	17.30	29.00	41.21	53.15	12.77	56.82
Nonlinear ridge-GMM with linear target (High-dim.)	28.18	27.72	2.75	16.86	30.71	45.26	58.42	12.91	45.78
Nonlinear ridge-GMM with dual target (High-dim.)	33.78	20.28	9.10	20.47	33.76	47.88	60.60	18.50	75.49
Nonlinear GMM with single synthetic assets (High-dim.)	32.97	27.41	5.54	20.92	36.13	50.69	63.12	17.66	63.09
Nonlinear GMM with dual synthetic assets (High-dim.)	38.01	20.82	11.71	24.40	38.80	53.04	64.98	22.70	94.21

Note: nonlinear ridge-GMM with shrinkage towards group targets is the estimator proposed in Koijen et al. (2023). Column $Diff(mean)$ provides the difference between the mean cross-out-of-sample R^2 of a given estimator and the Koijen et al. (2023) estimator. Column *t-stat* shows the t-stats associated with the test of the significance in the difference reported in column $Diff(mean)$. The table shows only the low-dimensional version of the nonlinear GMM with shrinkage towards group target since for high-dimensional demand functions, the estimator systematically does not converge. In line with Koijen et al. (2023), the demand functions are estimated for each fund annually. To mitigate the impact of outliers on the mean, I winsorize the distribution of R^2 for each estimator at 2.5%. Standard errors are clustered at mutual fund level.

$$R^2_{OOS,i,t}(M) = 1 - \frac{MSE_{OOS,i,t}(M)}{MSE_{OOS,i,t}(EW)}$$

Figure 6 illustrates the approach to the split of the data for the cross-out-of-sample exercise. Figure 7 shows the results of cross-out-of-sample validation: both across fund-years and in time series. Table 1 shows the details of the distribution of the cross-out-of-sample (COOS) R^2 for competing estimators. In column *t-stat*, I test the differences between means of the distributions of COOS R^2 for the estimator of Koijen et al. (2023) and the competing estimator in a given row.

The results of the cross-out-of-sample analysis overall suggest that demand estimation with data augmentation proposed in section 2 outperforms substantially the method of Koijen et al. (2023) both in the high-dimensional and low dimensional settings. The mean of the distribution of the cross-out-of-sample R^2 across fund-years is roughly twice larger for the low-dimensional (including only 8 baseline stock characteristics) nonlinear GMM with data augmentation compared to the low-dimensional ridge-GMM with shrinkage towards group targets proposed by Koijen et al. (2023). For high-dimensional demand functions – comprising 170 characteristics including characteristics from factor zoo and SIC 2-digit industry dummies – the ridge-GMM with

shrinkage towards group targets systematically doesn't converge. In fact, all alternative estimators in Table 1 have both economically and statistically higher mean cross-out-of-sample R^2 compared to the ridge-GMM with shrinkage towards group targets, suggesting that shrinking towards the group-level estimates masks important heterogeneity across mutual funds.

6 Application: Competition and Differentiation

Recently, there has been a rapid growth of the interest in the industrial organization of asset management. In this paper, I provide an evidence of a new channel of mutual fund differentiation from its peers. I show that in response to the increase in the competition for alpha, active mutual funds start to pursue investment strategies that are less similar to the strategies of their peers in a given style.

6.1 Asset demand function similarity

In the previous sections, I developed methodology that allows robust, high-quality estimation of the high-dimensional asset demand functions of individual institutional investors. Now, I propose to use the obtained estimates of the demand functions to define a measure of similarity of institutional investors in the investment strategy space. Specifically, define the *asset demand function similarity* between investors i and l at time t as:

$$CosineSim(\hat{\beta}_{i,t}, \hat{\beta}_{l,t}) = \frac{\left(\hat{\beta}_{i,t}\right)^T \hat{\beta}_{l,t}}{\|\hat{\beta}_{i,t}\|_2 \|\hat{\beta}_{l,t}\|_2}$$

Economically, this measure considers two institutional investors to be similar in terms of their investment strategy if their asset demand functions are similar. Unlike univariate measures based on investment-weighted characteristics (Hoberg et al. (2018), Lettau et al. (2018)), the cosine similarity between the high-dimensional demand function estimates has the following advantages: 1) it's multivariate; 2) it allows to study the implications of investor differentiation on asset prices through the structural model of Koijen and Yogo (2019), Koijen et al. (2023). Furthermore, contrary to the prospectus-

based measures of the mutual fund similarity (Kostovetsky and Warner (2020), Abis and Lines (2024)), my measure has the advantage of capturing the actual investment strategies pursued by the fund rather than measuring how mutual funds describe their strategies to their investors. Another advantage over prospectus-based measures is that holdings data are available not only for mutual funds but also for other institutional investors through the 13F fillings.

A special case of asset demand function similarity is the similarity between investor i and the average investor in a given style:

$$CosineSim(\hat{\beta}_{i,t}^{fund}, \hat{\beta}_{s(i),t}^{centro}) = \frac{\left(\hat{\beta}_{i,t}^{fund}\right)^T \hat{\beta}_{s(i),t}^{centro}}{\|\hat{\beta}_{i,t}^{fund}\|_2 \|\hat{\beta}_{s(i),t}^{centro}\|_2} \tag{24}$$

where $s(i)$ denotes that the style s in $\hat{\beta}^{centro}$ depends on the investor i. The demand function centroid $\beta_{s(i),t}^{centro}$ is defined as

$$\hat{\beta}_{s(i),t}^{centro} = \frac{1}{|\{l \in s(i)\}|} \sum_{l=1}^{|\{l \in s(i)\}|} \hat{\beta}_{l,t}^{fund} \tag{25}$$

where $|\{l \in s(i)\}|$ denotes the number investors l that belong to the same style $s(i)$ as investor i. In general, the measure in (24) does not put any restriction on the definition of the set of styles $\mathcal{S}$. The natural candidates for $\mathcal{S}$ in the mutual fund industry are Morningstar Mutual Fund Style Box, intransitive peer groups of Hoberg et al. (2018), CRSP Mutual Fund Objective Code.

6.2 Response of mutual funds to increase in competition

Framework and identification. Before presenting the identification strategy and empirical results, it is useful to outline the framework that will guide their interpretation. I start with a setup where mutual funds face diseconomies of scale on the capital they manage in the spirit of Berk and Green (2004), Barras et al. (2022):

$$\alpha_{i,t} = a_{i,t} - b_{i,t} q_{i,t-1} + \epsilon_{i,t} \tag{26}$$

where $\alpha_{i,t}$ is fund i's risk-adjusted performance at time t; $a_{i,t}$ is fund i's *skill* defined as the risk-adjusted performance on the first dollar, $q_{i,t-1}$ are asset under management (AUM) that fund i has at time $t-1$. The term $b_{i,t}q_{i,t-1}$ with $b_{i,t} > 0$ captures the diseconomies of scale at the fund level. That is, the more capital $q_{i,t-1}$ fund has under management, the lower is the risk-adjusted performance of fund i. The economic rationale behind such diseconomies are price impact and scarcity of investment ideas. The latter implies that as fund's size grows, it becomes harder to find very good investment ideas, and fund manager has to opt for the less promising investment ideas.

To study competition among funds for alpha, it is necessary to incorporate the effect of fund i's peers on fund i's performance. For this, I decompose fund i's alpha on the alpha of fund i's style $\alpha_{s(i),t}$ and the style-adjusted alpha $\alpha_{0,i,t}$:

$$\alpha_{i,t} = \alpha_{0,i,t} + \alpha_{s(i),t} = a_{0,i,t} - b_{0,i,t}q_{0,i,t-1} + a_{s(i),t} - b_{s(i),t}q_{s(i),t-1} + \epsilon_{i,t} + \epsilon_{s(i),t} \tag{27}$$

where the subscript $s(i)$ denotes the style s to which the fund i belongs to. The term $b_{s(i),t}q_{s(i),t-1}$, $b_{s(i),t} > 0$ corresponds to the diseconomies of scale at the style level. This term effectively captures how fund i's competitors – defined as funds that belong to the same style as fund i – impact fund i's risk-adjusted performance through the diseconomies of scale at the style level.[20]

In order to investigate the effect of competition among funds for alpha, one needs an exogenous shock to $q_{s(i),t-1}$. In an ideal experiment, an empiricist would like to compare the response of a group of funds that have been "randomly treated" by increase in competition (here, defined as competitors' AUM $q_{s(i),t-1}$) to the response of a control group that didn't experience such an increase. As an approximation of this ideal setting, I exploit the Morningstar mutual fund ratings reform in June 2002, which changed the way ratings are computed. Before June 2002, mutual ratings were assigned based on the metric of fund's performance that was *not* adjusted for the fund's style. This feature of the ratings methodology was thus disadvantaging those mutual funds which operated in poor-performing styles, while overestimating the quality of funds

[20]The latter can be interpreted as effect of crowding in trades (collective price impact of funds) or crowding in ideas (the more funds search for and implement investment strategies in a given style, the less unexploited investment ideas are available).

from well-performing styles (see Ben-David et al. (2020)). After June 2002, Morningstar updated its methodology to account for the funds' style using the well-know 3×3 Mutual Fund Style Box. Morningstar mutual fund ratings have been shown to direct fund flows (Ben-David et al., 2020), which motivates the relevance condition of style-level AUM being affected by the Morningstar reform.

Specifically, I perform 2SLS estimation of the following fund-quarter panel regression:

$$CosineSim(\hat{\beta}^{fund}_{i,t+h}, \hat{\beta}^{centro}_{s(i),t+h}) = \theta_h PctFlowCompet_{i,t} + Controls_{i,t} + FE_i + FE_t + \epsilon_{i,t+h} \tag{28}$$

where $CosineSim(\hat{\beta}^{fund}_{i,t+h}, \hat{\beta}^{centro}_{s(i),t+h})$ is the asset demand function similarity between fund i in quarter $t+h$ and the centroid of fund i's style in the same quarter $t+h$, defined in (25). Variable $PctFlowCompet_{i,t}$ is the quarterly percentage flow to fund i's competitors defined as other funds in the same style. Specifically,

$$PctFlowCompet_{i,t} = \frac{\sum\limits_{j \in s(i),\, j \neq i} DollarFlow_{j,t}}{\sum\limits_{j \in s(i),\, j \neq i} q_{j,t-1}} \times 100\% \tag{29}$$

Effectively, $PctFlowCompet_{i,t}$ represents the (percentage) change in $q_{s(i),t-1}$ from (27). The dollar flow and percentage flow of individual fund i are defined as follows:

$$DollarFlow_{i,t} = q_{j,t} - q_{j,t-1} \cdot (1 + r_{j,t}) \tag{30}$$

$$PctFlow_{i,t} = \frac{DollarFlow_{i,t}}{q_{j,t-1}} \times 100\% \tag{31}$$

I estimate equation (28) separately for each horizon h ranging from the first quarter to the quarter H after the reform. Control variables include logarithm of fund i's AUM, percentage flow to fund i, fund i's raw return at time t as well as AUM-weighted raw return of competitors. I also include fund and quarter fixed effects to absorb time-invariant unobserved heterogeneity across mutual funds and common shocks across time periods.

To account for endogeneity, $PctFlowCompet_{i,t}$ is instrumented with the style's

Table 2: First stage of 2SLS estimation

Dependent Variable:	PctFlowCompet					
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
ExposureMSshock x Post	1.513***	1.478***	1.508***	1.513***	1.478***	1.508***
	(0.0215)	(0.0230)	(0.0254)	(0.2351)	(0.2352)	(0.2593)
log(FundAUM)		-0.3135***	-0.3257***		-0.3135***	-0.3257***
		(0.0659)	(0.0674)		(0.0688)	(0.0552)
PctFlowFund		0.0079***	0.0083***		0.0079**	0.0083***
		(0.0016)	(0.0016)		(0.0029)	(0.0022)
ReturnFund		0.0132***	0.0158***		0.0132	0.0158***
		(0.0040)	(0.0041)		(0.0078)	(0.0044)
ReturnCompetVW			-0.0507***			-0.0507
			(0.0169)			(0.1311)
Fixed-effects						
Fund	Yes	Yes	Yes	Yes	Yes	Yes
Quarter	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics						
Cluster S.E.	Fund	Fund	Fund	Style	Style	Style
Observations	26,186	26,086	26,086	26,186	26,086	26,086
R^2	0.66544	0.66678	0.66724	0.66544	0.66678	0.66724
Within R^2	0.24960	0.25239	0.25341	0.24960	0.25239	0.25341

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Note: This table shows the results of the first stage of 2SLS estimation of (28). The percentage flow to fund i's competitors $PctFlowCompet_{i,t}$ is instrumented with the exposure to the Morningstar mutual fund ratings reform $ExposureMSshock_{s(i),t} \times Post_t$. Variable $Post_t$ is equal to 1 after the reform took place (2002 Q3 onwards), and zero otherwise.

$$PctFlowCompet_{i,t} = \kappa \cdot ExposureMSshock_{s(i),t} \times Post_t + Controls_{i,t} + FE_i + FE_t + \eta_{i,t}$$

Clustered standard errors are reported in parenthesis.

exposure to the Morningstar mutual fund ratings reform interacted with a dummy indicated that the reform took place, $ExposureMSshock_{s(i),t} \times Post_t$. Specifically, $Post_t$ is equal to 1 starting from the Q3 of 2002 – after the enactment of the reform in June 2002. To define the set of mutual fund styles $\mathcal{S}$, I use Morningstar 3×3 style box. The measure $ExposureMSshock_{s(i),t}$ is constructed as a rank ranging from -4 if the style was most negatively affected by the reform (largest average rating downgrade), to +4 if the style was most positively affected (largest average rating upgrade).[21] For the purpose of the construction of the instrument, the fund's style $s(i)$ is defined as the last style the fund i belong to just before the reform.

The exclusion restriction in (28) relies on the identifying assumption that the only channel through which Morningstar mutual fund ratings reform impacted funds' portfolio decisions are the diseconomies of scale at the style level. To account for the

[21] Since there are 9 styles, there are nine rank "levels" between -4 and +4.

potential impact of the Morningstar reform through the style performance, I control in my all specifications for the style-level performance of fund i's competitors.

The first stage of 2SLS estimation in (28) is strong with t-stat comfortably surpassing the threshold of 4.05 proposed by Stock and Yogo (2005). The results for the first stage are reported in Table 2.

Results. The results of OLS and 2SLS estimation of (28) across horizons up to 16 quarters are presented in Figure 8. The results of the second stage of 2SLS estimation are also shown in Table 3. The coefficient plot suggests that mutual funds start to differentiate from their competitors around 1 year after the increase in competition for alpha. Since the dependent variable $CosineSim(\hat{\beta}^{fund}_{i,t+h}, \hat{\beta}^{centro}_{s(i),t+h})$ is standardized,[22] the estimates suggest that 10% increase in flows to fund i's competitors lead to the decrease in cosine similarity by approximately 0.18 standard deviations, which is economically significant.

One can note the absence of a pre-trend in the coefficient plot of 2SLS in (8). In the context of my specification, it means that the part of variation in competitors' flows coming from the Morningstar mutual fund rating reform does not correlate with mutual fund differentiation *before* reform. This supports the identifying assumption by showing that the reform was not, on average, anticipated by active mutual funds.

While both OLS and 2SLS estimates provide qualitatively consistent conclusion about the decrease of the similarity between those funds exposed to the increase in competition and their peers, two distinctions are worth to note. First, the magnitude of 2SLS coefficient is about 3 times larger than that of OLS. One possible interpretation of this finding is that OLS estimation of (28) suffers from the measurement error. Implicit assumption in (28) is that we correctly measure which mutual funds are considered by a given fund i to be its competitors. It is possible that the actual set of competitors according to fund i's *subjective* measure of competition can be different from the 3×3 Morningstar style box used in my analysis. Alternatively, it could be that cross-diseconomies of scale[23] $\mathbb{E}_i[b_l]$, $l \in s(i)$ is heterogeneous according to fund i's

[22] by pre-June 2002 standard deviation.

[23] How fund j's AUM q_j affects fund i's ability to find and exploit investment opportunities

expectations, and thus, the correct measure of the competition should be a weighted by subjective proximity of competitors version of the flows to fund i's competitors. In the presence of measurement error in the independent variable, the OLS estimates in (28) can be biased towards zero. If Morningstar mutual fund ratings reform is uncorrelated with this measurement error, the 2SLS estimation would eliminate the above-mentioned attenuation bias, which can reconcile the notable difference between the magnitudes of OLS and 2SLS estimates.

Second notable difference between OLS and 2SLS estimates presented in Figure 8 is that mutual differentiation starts almost immediately according to the OLS results, but only after about 6-8 quarters according to the 2SLS estimates. One possibility is that mutual funds anticipate the increase in competition, and adjust their investment strategies before or at the same time as the competition is growing. Since OLS estimation of (28) is not robust to this type of endogeneity, one is likely to observe a much earlier onset of the differentiation. However, with instrumented competition, the expectations-driven endogeneity is mitigated since the Morningstar mutual fund reform was not largely anticipated by mutual funds (Ben-David et al. (2022)).

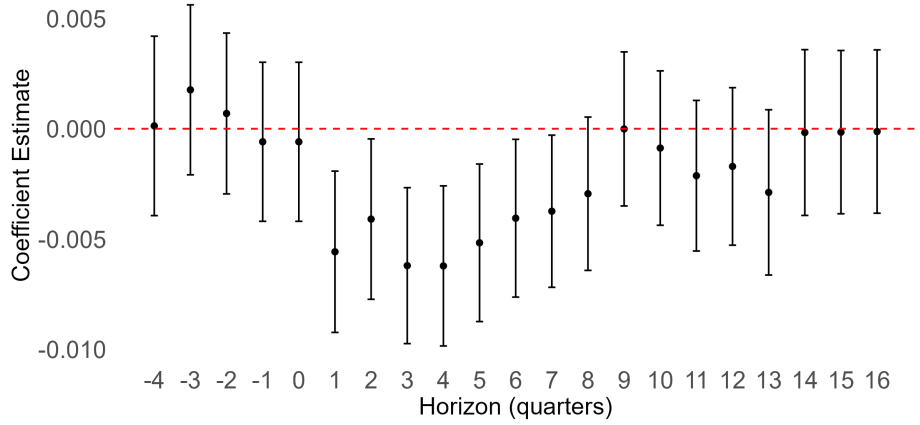
To show that the finding displayed in Figure 8 is not an artifact of estimation with synthetic assets, I replicate the same results using alternative high-dimensional estimators. As can be seen in Figure 11, the results are qualitatively and quantitatively similar if the demand function coefficients are estimated using alternative specification of synthetic assets.

Interestingly, the result on differentiation can not be obtained using univariate approach of investment-weighted stock characteristics used in Hoberg et al. (2018), Lettau et al. (2018). To show this, I construct the measure of similarity between the mutual fund i and the centroid of fund i's style $s(i)$ as the similarity between the two vectors of portfolio-weighted stock characteristics: $CosineSim(c^{fund}_{i,t+h}, c^{centro}_{s(i),t+h})$, where:

$$c^{fund}_{i,t} = \sum_{j=1}^{n_{i,t}} w_{i,j,t} c_{j,t} \tag{32}$$

where $c_{j,t} \in \mathbb{R}^{p\times 1}$ is a vector of p characteristics of stock j. To ensure the comparability

(a) OLS



(b) 2SLS

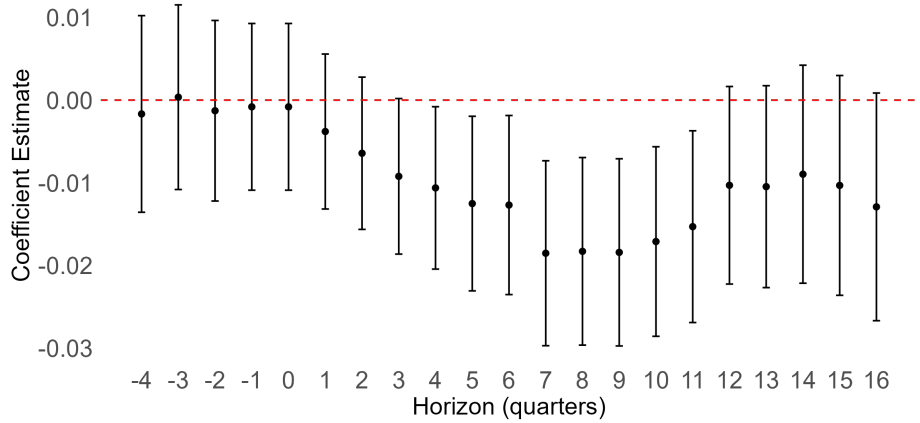


Figure 8: Mutual fund differentiation in response to increased competition

Note: This figure plots the point estimates and 95% confidence intervals for θ_h obtained by estimating (28) for $h = -4, ..., 16$. Panel A shows estimates obtained through OLS estimation of (28). Panel B presents the results of 2SLS estimation of (28), where the percentage flow to fund i's competitors $PctFlowCompet_{i,t}$ is instrumented with the exposure to the Morningstar mutual fund ratings reform $ExposureMSshock_{s(i),t} \times Post_t$. Controls include the logarithm of fund i's AUM, return of fund i, percentage flow of fund i as well as AUM-weighted returns of fund i's competitors. Cosine similarity between asset demand functions $CosineSim(\hat{\beta}^{fund}_{i,t+h}, \hat{\beta}^{centro}_{s(i),t+h})$ is based on nonlinear GMM estimation under moment conditions (14) with dual data augmentation. The dependent variable is normalized by the pre-June 2002 standard deviation. Standard errors are clustered at the fund level.

across characteristics, I rank-normalize all continuous stock characteristics to $[0, 1]$. The style centroid of fund i's style is defined analogously to (25), with funds' $\beta \in \mathbb{R}^{p\times 1}$ being replaced by funds' characteristic vector $c \in \mathbb{R}^{p\times 1}$.

As can be seen in Figure 9, both OLS and 2SLS deliver null results. This supports the usefulness of the high-dimensional multivariate approach to the measurement of

Table 3: Second stage of 2SLS estimation

Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Variables												
PctFlowCompet	-0.0038	-0.0064	-0.0092*	-0.0106**	-0.0125**	-0.0127**	-0.0185***	-0.0183***	-0.0184***	-0.0171***	-0.0153***	-0.0103*
	(0.0048)	(0.0047)	(0.0048)	(0.0050)	(0.0054)	(0.0055)	(0.0057)	(0.0058)	(0.0058)	(0.0058)	(0.0059)	(0.0061)
log(FundAUM)	-0.0029	-0.0018	-0.0027	0.0012	0.0198	0.0368	0.0346	0.0244	0.0329	0.0125	-0.0003	0.0097
	(0.0213)	(0.0217)	(0.0222)	(0.0232)	(0.0248)	(0.0259)	(0.0265)	(0.0283)	(0.0294)	(0.0304)	(0.0310)	(0.0311)
PctFlowFund	-0.0007	-0.0005	-0.0008*	-0.0005	-0.0004	-0.0002	0.0003	0.0005	0.0009*	0.0007	0.0010**	0.0002
	(0.0004)	(0.0004)	(0.0004)	(0.0005)	(0.0005)	(0.0005)	(0.0005)	(0.0005)	(0.0005)	(0.0005)	(0.0004)	(0.0005)
ReturnFund	-0.0003	0.0001	-0.0010	-0.0014*	-0.0011	-0.0007	0.0004	-0.0006	-0.00008	-0.0004	-0.0015**	0.0005
	(0.0008)	(0.0007)	(0.0008)	(0.0008)	(0.0008)	(0.0007)	(0.0008)	(0.0008)	(0.0008)	(0.0007)	(0.0007)	(0.0008)
ReturnCompetVW	-0.0009	-0.0014	0.0039*	-0.0015	-0.00008	-0.0026	0.0034	0.0007	0.0024	0.0020	0.0020	-0.0008
	(0.0024)	(0.0024)	(0.0023)	(0.0022)	(0.0021)	(0.0023)	(0.0024)	(0.0026)	(0.0025)	(0.0024)	(0.0026)	(0.0025)
Fixed-effects												
Fund	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Quarter	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics												
Cluster S.E.	Fund	Fund	Fund	Fund	Fund	Fund	Fund	Fund	Fund	Fund	Fund	Fund
Observations	23,596	22,603	21,626	20,667	19,719	18,776	17,850	16,918	15,984	15,036	14,093	13,154
R^2	0.3616	0.3683	0.3762	0.3812	0.3863	0.3904	0.3917	0.3965	0.4068	0.4160	0.4319	0.4481
Within R^2	0.0007	0.0004	0.0011	0.0010	0.0001	-0.0001	-0.0034	-0.0042	-0.0067	-0.0057	-0.0032	-0.0016

Clustered (Fund) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Note: This table presents the results of 2SLS estimation of (28), where the percentage flow to fund i's competitors $PctFlowCompet_{i,t}$ is instrumented with the exposure to the Morningstar mutual fund ratings reform $ExposureMSshock_{s(i),t} \times Post_t$. Each column corresponds to the horizon $h \in 1, ...12$ at which the effect of increased competition on differentiation is estimated. Controls include the logarithm of fund i's AUM, return of fund i, percentage flow of fund i as well as AUM-weighted returns of fund i's competitors. Cosine similarity between asset demand functions $CosineSim(\hat{\beta}^{fund}_{i,t+h}, \hat{\beta}^{centro}_{s(i),t+h})$ is based on nonlinear GMM estimation under moment conditions (14) with dual data augmentation. The dependent variable is normalized by the pre-June 2002 standard deviation. Standard errors are clustered at the fund level and are reported in parenthesis.

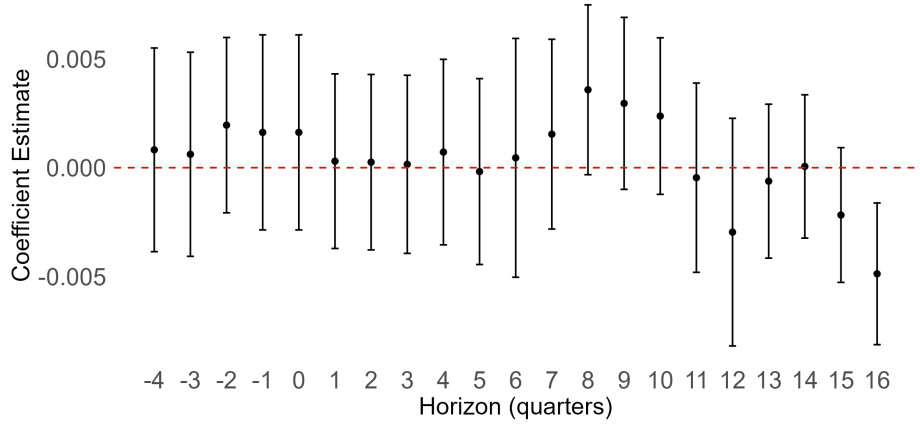
investment strategies from mutual fund holdings.

Overall, the results suggest that active mutual funds respond to the increase in competition through differentiation. The economic implications of this finding are two-fold. First, increase in competition encourages innovation in the investment strategy space through the search of investment opportunities that are different from those exploited by competitors, broadly consistent with the theoretical work on escape competition effect Aghion et al. (2005), Aghion et al. (2009). Second, as mutual funds differentiate within the style, their risk exposures are likely to change as well. This provides support for the customized-peer performance evaluation, such as in Hoberg et al. (2018), Abis and Lines (2024).

6.3 Channels of differentiation

Another advantage of using high-dimensional demand functions to measure differentiation between funds is that it is straightforward to compute partial similarity with respect

(a) OLS



(b) 2SLS

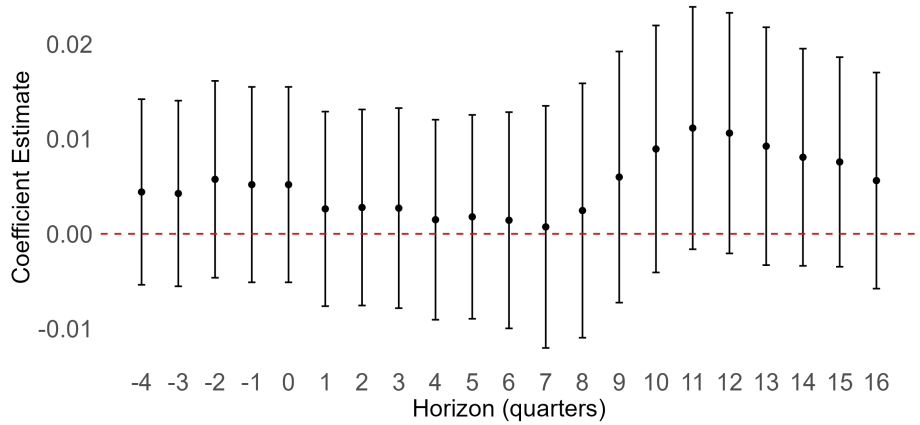


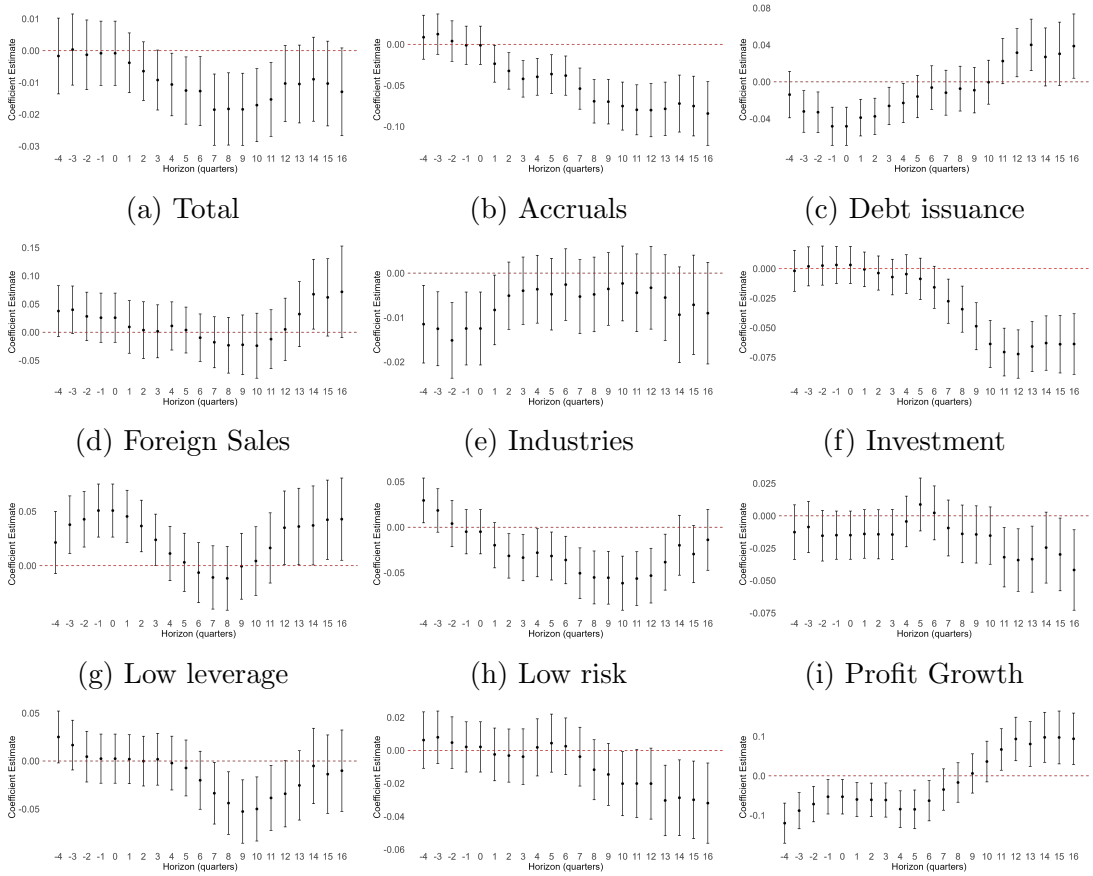
Figure 9: Mutual fund differentiation in response to increased competition: 2SLS, naïve estimators

Note:

Controls include the logarithm of fund i's AUM, return of fund i, percentage flow of fund i as well as AUM-weighted returns of fund i's competitors. Cosine similarity between asset demand functions $CosineSim(\hat{\beta}^{fund}_{i,t+h}, \hat{\beta}^{centro}_{s(i),t+h})$ is based on nonlinear GMM estimation under moment conditions (14) with dual data augmentation as proposed in section 2. The dependent variable is normalized by the pre-June 2002 standard deviation. Standard errors are clustered at the fund level.

to a given set of stock characteristics. For example, given the evidence of mutual fund differentiation in response to increased competition, one might wonder *which* demand function loadings drive this differentiation. In this subsection, I further explore which themes of stock characteristics (defined following Jensen et al. (2023)) do mutual funds differentiate along.

I define partial cosine similarity between asset demand functions as the cosine



(a) Total (b) Accruals (c) Debt issuance

(d) Foreign Sales (e) Industries (f) Investment

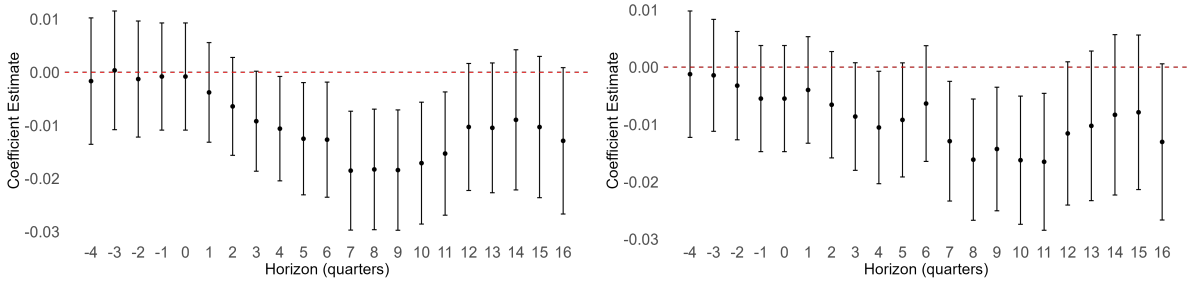
(g) Low leverage (h) Low risk (i) Profit Growth

(j) Profitability (k) Quality (l) Size

Figure 10: Partial differentiation along subsets of stock characteristics

Note: The dependent variable in all panels is normalized by the pre-June 2002 standard deviation of the *total* cosine similarity (to make the magnitude comparable across all panels). Standard errors are clustered by fund.

similarity between loadings of demand functions specific to a given subset of stock characteristics. For example, a partial cosine similarity with respect to profitability theme is the cosine similarity between the two vectors of demand function loadings on stock characteristics comprising the profitability theme. I plot the results of 2SLS estimation for each of the characteristic themes as well for the industry dummies in Figure 10. Results suggest that mutual funds differentiate predominantly along the exposures to characteristics in investment, accruals, and profitability theme, but not through the differentiated industry exposures.



(a) nonlinear GMM with (dual) synthetic assets

(b) nonlinear GMM with (single) synthetic assets

Figure 11: Robustness to the choice of the demand function estimator: Mutual fund differentiation in response to increased competition
Note: The dependent variable is normalized by the pre-June 2002 standard deviation. Standard errors are clustered by fund.

7 Robustness

7.1 Alternative high-dimensional multivariate estimators

In this subsection, I replicate the analysis of the effect of increased competition on mutual fund differentiation using alternative high-dimensional estimators. The results the displayed in Figure 11. The main result is robust to the specification of synthetic assets.

8 Conclusion

In this paper, I address the issue that many institutional investors hold concentrated portfolios, which makes it challenging to thoroughly describe the individual demand of institutional investors. I propose a data augmentation technique based on the generation of data-driven and economically interpretable synthetic assets. I show that this data augmentation acts as an adaptive shrinkage which automatically adjusts the shrinkage rate to the cost of overfitting faced by the nonlinear demand function estimator. The resulting estimation technique leads to substantial improvement in cross-out-of-sample R^2 for estimation of both low-dimensional and high-dimensional demand functions.

I use the proposed methodology to construct a measure of investor differentiation. Using the Morningstar mutual fund ratings reform in 2002 as a shock to competition for alpha, I show that mutual funds escape the increased competition through differen-

tiation from the competitors. The economic implications of this finding are two-fold. First, increase in competition encourages innovation in the investment strategy space through the search of investment opportunities that are different from those exploited by competitors, broadly consistent with the theoretical work on escape competition effect Aghion et al. (2005), Aghion et al. (2009). Second, as mutual funds differentiate within the style, their risk exposures are likely to change as well. This provides support for the customized-peer performance evaluation, such as in Hoberg et al. (2018), Abis and Lines (2024).

References

Abis, S. (2020). Man vs. Machine: Quantitative and Discretionary Equity Management.

Abis, S. and Lines, A. (2024). Broken promises, competition, and capital allocation in the mutual fund industry. *Journal of Financial Economics*, 162:103948.

Aghion, P., Bloom, N., Blundell, R., Griffith, R., and Howitt, P. (2005). Competition and Innovation: an Inverted-U Relationship*. *The Quarterly Journal of Economics*, 120(2):701–728.

Aghion, P., Blundell, R., Griffith, R., Howitt, P., and Prantl, S. (2009). The Effects of Entry on Incumbent Innovation and Productivity. *The Review of Economics and Statistics*, 91(1):20–32.

Barras, L., Gagliardini, P., and Scaillet, O. (2022). Skill, Scale, and Value Creation in the Mutual Fund Industry. *The Journal of Finance*, 77(1):601–638.

Ben-David, I., Li, J., Rossi, A., and Song, Y. (2020). Non-Fundamental Demand and Style Returns. *SSRN Electronic Journal*.

Ben-David, I., Li, J., Rossi, A., and Song, Y. (2022). Ratings-Driven Demand and Systematic Price Fluctuations. *The Review of Financial Studies*, 35(6):2790–2838.

Berk, J. and Green, R. (2004). Mutual Fund Flows and Performance in Rational Markets. *Journal of Political Economy*, 112(6):1269–1295. Publisher: The University of Chicago Press.

Bonelli, M. (2022). Data-driven Investors.

Bonelli, M., Buyalskaya, A., and Yao, T. (2021). Quality and Product Differentiation: Theory and Evidence from the Mutual Fund Industry. *SSRN Electronic Journal*.

Bonelli, M. and Foucault, T. (2023). Displaced by Big Data: Evidence from Active Fund Managers.

Bretscher, L., Schmid, L., Sen, I., and Sharma, V. (2022). Institutional Corporate Bond Pricing.

Bryzgalova, S., Lerner, S., Lettau, M., and Pelger, M. (2024). Missing Financial Data. *The Review of Financial Studies*, page hhae036.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Cong, L. W., Tang, K., Wang, J., and Zhang, Y. (2021). AlphaPortfolio: Direct Construction Through Deep Reinforcement Learning and Interpretable AI.

de Jong, S. (1993). SIMPLS: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18(3):251–263.

Didisheim, A., Ke, S., Kelly, B. T., and Malamud, S. (2023). APT or "AIPT"? The Surprising Dominance of Large Factor Models.

Dohmatob, E., Feng, Y., and Kempe, J. (2024). Model Collapse Demystified: The Case of Regression. arXiv:2402.07712.

Dugast, J. and Foucault, T. (2023). Equilibrium Data Mining and Data Abundance.

Freyberger, J., Hoeppner, B., Neuhierl, A., and Weber, M. (2024). Missing Data in Asset Pricing Panels. *The Review of Financial Studies*, page hhae003.

Gabaix, X., Koijen, R. S. J., Richmond, R., and Yogo, M. (2024). Asset Embeddings.

Giglio, S., Liao, Y., and Xiu, D. (2021). Thousands of Alpha Tests. *The Review of Financial Studies*, 34(7):3456–3496.

Gu, S., Kelly, B., and Xiu, D. (2020). Empirical Asset Pricing via Machine Learning. *The Review of Financial Studies*, 33(5):2223–2273.

Haddad, V., Huebner, P., and Loualiche, E. (2021). How Competitive is the Stock Market? Theory, Evidence from Portfolios, and Implications for the Rise of Passive Investing. *SSRN Electronic Journal*.

Hoberg, G., Kumar, N., and Prabhala, N. (2018). Mutual Fund Competition, Managerial Skill, and Alpha Persistence. *The Review of Financial Studies*, 31(5):1896–1929.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge Regression: Applications to Nonorthogonal Problems. *Technometrics*, 12(1):69–82. Publisher: [Taylor & Francis, Ltd., American Statistical Association, American Society for Quality].

Hommel, N., Landier, A., and Thesmar, D. (2021). Corporate Valuation: An Empirical Comparison of Discounting Methods.

Huang, D., Stein, N., Rubin, D. B., and Kou, S. C. (2020). Catalytic prior distributions with application to generalized linear models. *Proceedings of the National Academy of Sciences of the United States of America*, 117(22):12004–12010.

Huebner, P. (2023). The Making of Momentum: A Demand-System Perspective. *SSRN Electronic Journal*.

Jensen, T. I., Kelly, B., and Pedersen, L. H. (2023). Is There a Replication Crisis in Finance? *The Journal of Finance*, 78(5):2465–2518. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jofi.13249.

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

Kelly, B., Malamud, S., and Zhou, K. (2024). The Virtue of Complexity in Return Prediction. *The Journal of Finance*, 79(1):459–503. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jofi.13298.

Kelly, B. T., Malamud, S., and Zhou, K. (2022). The Virtue of Complexity Everywhere.

Koijen, R. S. J., Koulischer, F., Nguyen, B., and Yogo, M. (2021). Inspecting the mechanism of quantitative easing in the euro area. *Journal of Financial Economics*, 140(1):1–20.

Koijen, R. S. J., Richmond, R. J., and Yogo, M. (2023). Which Investors Matter for Equity Valuations and Expected Returns? *The Review of Economic Studies*, page rdad083.

Koijen, R. S. J. and Yogo, M. (2019). A Demand System Approach to Asset Pricing. *Journal of Political Economy*, 127(4):1475–1515. Publisher: The University of Chicago Press.

Koijen, R. S. J. and Yogo, M. (2024). Exchange Rates and Asset Prices in a Global Demand System.

Kostovetsky, L. and Warner, J. B. (2020). Measuring Innovation and Product Differentiation: Evidence from Mutual Funds. *The Journal of Finance*, 75(2):779–823. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jofi.12853.

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

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Lettau, M., Ludvigson, S. C., and Manoel, P. (2018). Characteristics of Mutual Fund Portfolios: Where Are the Value Funds?

Li, S. and Qiu, J. (2014). Financial Product Differentiation over the State Space in the Mutual Fund Industry. *Management Science*, 60(2):508–520. Publisher: INFORMS.

Li, Y. and Liu, F. (2022). Adaptive Noisy Data Augmentation for Regularized Estimation and Inference in Generalized Linear Models.

Martin, I. W. R. and Nagel, S. (2022). Market efficiency in the age of big data. *Journal of Financial Economics*, 145(1):154–177.

Nenova, T. (2024). Global or Regional Safe Assets: Evidence from Bond Substitution Patterns.

Noh, D., Oh, S., and Song, J. (2020). Unpacking the Demand for Sustainable Equity Investing.

Plazzi, A., Tamoni, A., and Zanotti, M. (2023). Financial Intermediaries and Demand for Duration.

Stock, J. H. and Yogo, M. (2005). Testing for Weak Instruments in Linear IV Regression. *2005*, page 48.

van der Beck, P. (2021). Flow-Driven ESG Returns.

van der Beck, P. (2022). On the Estimation of Demand-Based Asset Pricing Models.

van Wieringen, W. N. (2023). Lecture notes on ridge regression. arXiv:1509.09169 [stat].

Wahal, S. and Wang, A. Y. (2011). Competition among mutual funds. *Journal of Financial Economics*, 99(1):40–59.

Zhu, Q. (2020). The Missing New Funds. *Management Science*, 66(3):1193–1204.

Zou, H. and Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2):301–320. Publisher: [Royal Statistical Society, Wiley].

Zou, H. and Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics*, 37(4). arXiv:0908.1836 [math, stat].

A Descriptive Statistics

A.1 Active Equity Mutual Funds

Table A.1: Number of stocks in portfolios of active equity mutual funds

Panel A: Post-2010, Quarterly

variable	mean	sd	min	p01	p10	p15	p20	p25	p50	p75	p90	p99	max
n stocks (w > 0)	183.2	311.4	13.0	21.0	32.0	38.0	44.0	52.0	103.0	185.0	332.0	1625.0	4692.0
n stocks (IU)	350.4	484.3	16.0	29.0	56.0	69.0	84.0	99.0	201.0	396.0	733.0	2488.6	5647.0

Panel B: Pre-2010, Quarterly

variable	mean	sd	min	p01	p10	p15	p20	p25	p50	p75	p90	p99	max
n stocks (w > 0)	80.9	92.6	19.0	21.0	30.0	34.0	37.0	40.0	57.0	87.0	142.0	437.0	2366.0
n stocks (IU)	172.4	156.6	20.0	33.0	57.0	66.0	75.0	84.0	129.0	205.0	322.0	806.0	3048.0

Panel C: Full Sample, Quarterly

variable	mean	sd	min	p01	p10	p15	p20	p25	p50	p75	p90	p99	max
n stocks (w > 0)	134.7	240.3	13.0	21.0	31.0	35.0	40.0	44.0	72.0	136.0	244.0	1209.5	4692.0
n stocks (IU)	266.1	378.1	16.0	30.0	56.0	67.0	78.0	89.0	154.0	293.0	540.0	1933.2	5647.0

Panel A: Post-2010, Annually

variable	mean	sd	min	p01	p10	p15	p20	p25	p50	p75	p90	p99	max
n stocks (w > 0)	649.5	1135.3	17.0	25.0	94.0	117.0	141.0	167.0	359.0	675.0	1191.0	6188.0	17162.0
n stocks (IU)	1242.1	1780.4	20.0	43.0	159.0	210.0	260.0	317.0	684.0	1421.0	2655.0	9026.1	21997.0

Panel B: Pre-2010, Annually

variable	mean	sd	min	p01	p10	p15	p20	p25	p50	p75	p90	p99	max
n stocks (w > 0)	240.8	299.5	19.0	23.0	61.0	77.0	91.0	104.0	170.0	275.0	444.0	1431.0	7028.0
n stocks (IU)	513.6	525.4	20.0	40.0	116.0	149.0	178.0	206.0	371.0	632.0	1040.5	2663.1	7563.0

Panel C: Full Sample, Annually

variable	mean	sd	min	p01	p10	p15	p20	p25	p50	p75	p90	p99	max
n stocks (w > 0)	438.2	842.9	17.0	24.0	72.0	92.0	107.0	124.0	224.0	446.0	833.0	4261.6	17162.0
n stocks (IU)	865.4	1343.9	20.0	41.0	133.0	170.0	206.0	243.0	472.0	953.0	1809.0	7079.7	21997.0

Note: In the table, "n stocks ($w > 0$)" correspond to the number of stocks actually held by mutual fund with strictly positive weight in fund's portfolio. The variable "n stocks (IU)" is the number of stocks in fund's investment universe, which is defined following Koijen and Yogo (2019), Koijen et al. (2023) as the set of stocks that have been held by a given fund over the last 12 quarters. The set of stocks includes *inside* assets as per definition of Koijen and Yogo (2019), Koijen et al. (2023), which are the assets on which the estimation of the demand function is performed. If a mutual fund does not report any holdings in a given year, this fund-year observation is not counted in the computation of descriptive statistics.

A.2 13F Institutional Investors

Table A.2 reports the descriptive statistics for the number of stocks held by institutional investors that report to 13F.

Table A.2: Number of stocks in portfolios of 13F institutional investors

Panel A: Post-2010, Quarterly

variable	mean	sd	min	p01	p10	p15	p20	p25	p50	p75	p90	p99	max
n stocks (w > 0)	229.8	343.7	20.0	21.0	34.0	41.0	48.0	54.0	97.0	224.0	582.0	1801.0	2510.0
n stocks (IU)	354.8	436.9	20.0	31.0	58.0	69.0	79.0	89.0	170.0	419.0	959.0	1989.0	2562.0

Panel B: Pre-2010, Quarterly

variable	mean	sd	min	p01	p10	p15	p20	p25	p50	p75	p90	p99	max
n stocks (w > 0)	244.6	388.0	20.0	22.0	38.0	45.0	51.0	58.0	102.0	237.0	583.0	2069.6	3517.0
n stocks (IU)	393.2	485.7	20.0	38.0	72.0	84.0	96.0	108.0	197.0	462.0	985.0	2365.0	3562.0

Panel C: Full Sample, Quarterly

variable	mean	sd	min	p01	p10	p15	p20	p25	p50	p75	p90	p99	max
n stocks (w > 0)	236.4	364.1	20.0	22.0	36.0	43.0	49.0	56.0	100.0	230.0	582.0	1895.0	3517.0
n stocks (IU)	371.8	459.5	20.0	33.0	63.0	75.0	86.0	97.0	182.0	440.0	970.0	2129.0	3562.0

Panel D: Post-2010, Annually

variable	mean	sd	min	p01	p10	p15	p20	p25	p50	p75	p90	p99	max
n stocks (w > 0)	801.1	1291.4	20.0	22.0	86.0	112.0	139.0	167.0	328.0	748.0	2037.0	7032.0	9936.0
n stocks (IU)	1236.6	1656.7	20.0	36.0	142.0	190.0	235.0	281.0	559.0	1408.0	3355.0	7757.0	10198.0

Panel E: Pre-2010, Annually

variable	mean	sd	min	p01	p10	p15	p20	p25	p50	p75	p90	p99	max
n stocks (w > 0)	840.2	1434.5	20.0	26.0	95.0	124.0	151.0	177.0	340.0	786.0	1996.0	7870.8	13768.0
n stocks (IU)	1350.8	1821.4	20.0	51.0	174.0	227.8	280.0	329.0	636.0	1554.0	3463.0	9157.0	13991.0

Panel F: Full Sample, Annually

variable	mean	sd	min	p01	p10	p15	p20	p25	p50	p75	p90	p99	max
n stocks (w > 0)	818.5	1357.3	20.0	24.0	89.0	117.0	145.0	171.0	333.0	765.0	2023.4	7276.9	13768.0
n stocks (IU)	1287.6	1733.1	20.0	42.0	157.0	205.0	253.0	300.0	594.0	1469.0	3408.0	8202.0	13991.0

Note: In the table, "n stocks ($w > 0$)" correspond to the number of stocks actually held by mutual fund with strictly positive weight in fund's portfolio. The variable "n stocks (IU)" is the number of stocks in fund's investment universe, which is defined following Koijen and Yogo (2019), Koijen et al. (2023) as the set of stocks that have been held by a given fund over the last 12 quarters. The set of stocks includes *inside* assets as per definition of Koijen and Yogo (2019), Koijen et al. (2023), which are the assets on which the estimation of the demand function is performed. If a 13F institutional investor does not report any holdings in a given year, this investor-year observation is not counted in the computation of descriptive statistics.

A.3 Passive Equity Mutual Funds

Table A.3: Number of stocks in portfolios of passive equity mutual funds

Panel A: Post-2010, Quarterly

variable	mean	sd	min	p01	p10	p15	p20	p25	p50	p75	p90	p99	max
n stocks (w > 0)	897.8	1018.1	18.0	28.0	119.0	158.0	214.0	248.0	539.0	1178.0	2035.9	5337.9	6401.0
n stocks (IU)	1181.0	1107.1	21.0	39.0	225.0	298.0	368.0	417.0	910.0	1455.0	2498.9	5644.8	6992.0

Panel B: Pre-2010, Quarterly

variable	mean	sd	min	p01	p10	p15	p20	p25	p50	p75	p90	p99	max
n stocks (w > 0)	467.8	459.4	19.0	25.0	72.0	90.0	133.0	177.0	410.0	458.0	1171.0	2180.5	3998.0
n stocks (IU)	580.7	504.9	20.0	31.0	119.0	204.0	263.0	317.0	435.0	572.0	1392.4	2285.0	4497.0

Panel C: Full Sample, Quarterly

variable	mean	sd	min	p01	p10	p15	p20	p25	p50	p75	p90	p99	max
n stocks (w > 0)	782.9	923.2	18.0	27.0	96.0	138.0	186.0	234.0	426.0	1079.0	1718.5	5155.6	6401.0
n stocks (IU)	1020.7	1018.3	20.0	35.0	198.0	271.0	340.0	389.0	666.0	1315.0	2144.0	5464.5	6992.0

Panel A: Post-2010, Annually

variable	mean	sd	min	p01	p10	p15	p20	p25	p50	p75	p90	p99	max
n stocks (w > 0)	3207.6	3846.4	20.0	58.0	327.0	461.0	604.0	803.0	1797.0	4313.0	7085.0	20800.8	25410.0
n stocks (IU)	4219.5	4220.2	27.0	97.0	546.0	857.0	1125.0	1368.0	3123.0	5370.0	8990.0	21858.4	25745.0

Panel B: Pre-2010, Annually

variable	mean	sd	min	p01	p10	p15	p20	p25	p50	p75	p90	p99	max
n stocks (w > 0)	1419.0	1517.8	20.0	44.6	164.0	241.9	337.6	419.0	1010.5	1685.2	3394.1	7549.5	10199.0
n stocks (IU)	1761.4	1693.3	20.0	58.0	324.3	424.0	523.2	721.2	1328.0	1828.5	4240.3	8243.1	11351.0

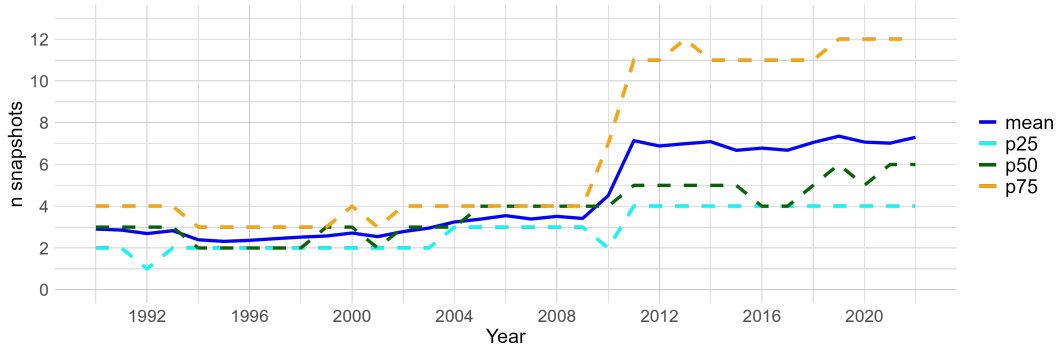
Panel C: Full Sample, Annually

variable	mean	sd	min	p01	p10	p15	p20	p25	p50	p75	p90	p99	max
n stocks (w > 0)	2670.5	3422.7	20.0	54.0	260.0	375.0	492.8	653.0	1468.0	3522.0	6022.6	19006.4	25410.0
n stocks (IU)	3481.3	3819.9	20.0	78.0	418.4	640.6	872.0	1065.0	2078.0	4736.0	7420.6	20240.0	25745.0

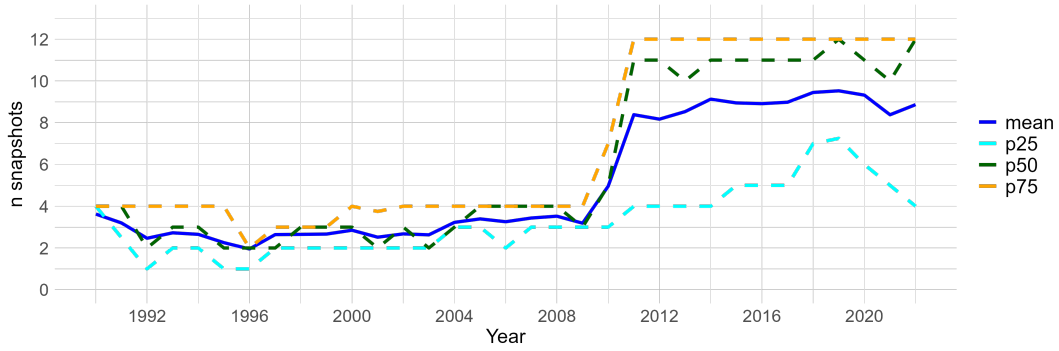
Note: In the table, "n stocks ($w > 0$)" correspond to the number of stocks actually held by mutual fund with strictly positive weight in fund's portfolio. The variable "n stocks (IU)" is the number of stocks in fund's investment universe, which is defined following Koijen and Yogo (2019), Koijen et al. (2023) as the set of stocks that have been held by a given fund over the last 12 quarters. The set of stocks includes *inside* assets as per definition of Koijen and Yogo (2019), Koijen et al. (2023), which are the assets on which the estimation of the demand function is performed. If a mutual fund does not report any holdings in a given year, this fund-year observation is not counted in the computation of descriptive statistics.

A.4 Frequency of Reporting by Institutional Investors

Panel A: Active equity mutual funds, Annually



Panel B: Passive equity mutual funds, Annually



Panel C: 13F institutions investors, Annually

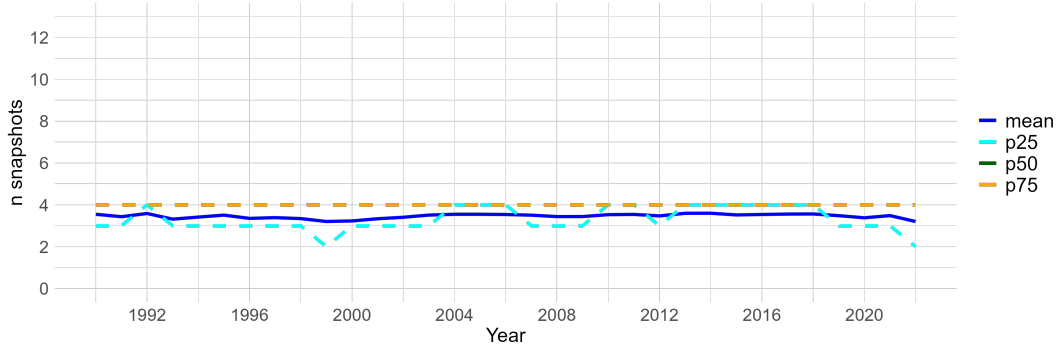


Figure A.1: Number of holdings' snapshots reported by institutional investors per year

Note: If a mutual fund (13F institutional investor) does not report any holdings in a given year, this fund-year (investor-year) observation is not counted in the computation of mean and percentiles.

A.5 List of Stock Characteristics

Table A.4: List of stock characteristics in high-dimensional specification

Code	Name	Theme	Rescaling	KY19 or KRY23 baseline
at_gr1	Asset Growth	Investment		Yes
beta_60m	Market beta	Low risk		Yes
ope_be	Operating profits-to-book equity	Profitability		Yes
lerner	Profit margin	Profitability		Yes
be	Book equity	Size		Yes
div12m_be	Dividend yield	Value	me	Yes
sale_be	Sales-to-market	Value	me	Yes
foreign_sales	Foreign sales	Foreign sales		Yes
cowc_gr1a	Change in current operating working capital	Accruals		-
oaccruals_at	Operating accruals	Accruals		-
oaccruals_ni	Percent operating accruals	Accruals		-
taccruals_ni	Percent total accruals	Accruals		-
taccruals_at	Total accruals	Accruals		-
fnl_gr1a	Change in financial liabilities	Debt issuance		-
ncol_gr1a	Change in noncurrent operating liabilities	Debt issuance		-
debt_gr3	Growth in book debt (3 years)	Debt issuance		-
dbnetis_at	Net debt issuance	Debt issuance		-
capex_abn	Abnormal corporate investment	Investment		-
capx_gr1	CAPEX growth (1 year)	Investment		-
capx_gr2	CAPEX growth (2 years)	Investment		-
capx_gr3	CAPEX growth (3 years)	Investment		-
ppeinv_gr1a	Change PPE and Inventory	Investment		-
be_gr1a	Change in common equity	Investment		-
coa_gr1a	Change in current operating assets	Investment		-
col_gr1a	Change in current operating liabilities	Investment		-
lti_gr1a	Change in long-term investments	Investment		-
lnoa_gr1a	Change in long-term net operating assets	Investment		-
nfna_gr1a	Change in net financial assets	Investment		-
nncoa_gr1a	Change in net noncurrent operating assets	Investment		-
noa_gr1a	Change in net operating assets	Investment		-
ncoa_gr1a	Change in noncurrent operating assets	Investment		-
sti_gr1a	Change in short-term investments	Investment		-
emp_gr1	Hiring rate	Investment		-
inv_gr1a	Inventory change	Investment		-
inv_gr1	Inventory growth	Investment		-
saleq_gr1	Sales Growth (1 quarter)	Investment		-
sale_gr1	Sales Growth (1 year)	Investment		-
sale_gr3	Sales Growth (3 years)	Investment		-
at_be	Book leverage	Low leverage		-
cash_at	Cash-to-assets	Low leverage		-
age	Firm age	Low leverage		-
netdebt_be	Net debt-to-price	Low leverage	me	-
ocfq_saleq_std	Cash flow volatility	Low risk		-
betadown_252d	Downside beta	Low risk		-
earnings_variability	Earnings variability	Low risk		-
betabab_1260d	Frazzini-Pedersen market beta	Low risk		-
ocf_at_chg1	Change in operating cash flow to assets	Profit Growth		-
niq_at_chg1	Change in quarterly return on assets	Profit Growth		-
niq_be_chg1	Change in quarterly return on equity	Profit Growth		-
dsale_dinv	Change sales minus change Inventory	Profit Growth		-

Continued on the next page.

List of stock characteristics in high-dimensional specification (Continued)

Code	Name	Theme	Rescaling	KY19 or KRY23 baseline
dsale_dsga	Change sales minus change SG&A	Profit Growth		-
dsale_drec	Change sales minus change receivables	Profit Growth		-
saleq_su	Standardized Revenue surprise	Profit Growth		-
niq_su	Standardized earnings surprise	Profit Growth		-
tax_gr1a	Tax expense surprise	Profit Growth		-
ocf_at	Operating cash flow to assets	Profitability		-
niq_be	Quarterly return on equity	Profitability		-
ni_be	Return on equity	Profitability		-
ebit_be	Return on net operating assets	Profitability	bev	-
pi_nix	Taxable income-to-book income	Profitability		-
at_turnover	Capital turnover	Quality		-
cop_atl1	Cash-based operating profits-to-lagged book assets	Quality		-
dgp_dsale	Change gross margin minus change sales	Quality		-
ni_ar1	Earnings persistence	Quality		-
gp_at	Gross profits-to-assets	Quality		-
sale_emp_gr1	Labor force efficiency	Quality		-
aliq_at	Liquidity of book assets	Quality		-
noa_at	Net operating assets	Quality		-
ni_inc8q	Number of consecutive quarters with earnings increases	Quality		-
opex_at	Operating leverage	Quality		-
op_at	Operating profits-to-book assets	Quality		-
niq_at	Quarterly return on assets	Quality		-
rd5_at	R&D capital-to-book assets	Quality		-
rd_be	R&D-to-market	Quality	me	-
rd_sale	R&D-to-sales	Quality		-
debt_be	Debt-to-market	Value	me	-
ebitda_be	Ebitda-to-market enterprise value	Value	mev	-
eqnpo_12m	Equity net payout	Value		-
fcf_be	Free cash flow-to-price	Value	me	-
eqnetis_at	Net equity issuance	Value		-
eqnpo_be	Net payout yield	Value	me	-
chcsho_12m	Net stock issues	Value		-
netis_at	Net total issuance	Value		-
ocf_be	Operating cash flow-to-market	Value	me	-
eqpo_be	Payout yield	Value	me	-

Note: Column *Code* provides the code name of the characteristic used within this study. The code names already incorporate the re-scaling by *be* if the latter is necessary.

A.6 Correlation Matrix of Stock Characteristics

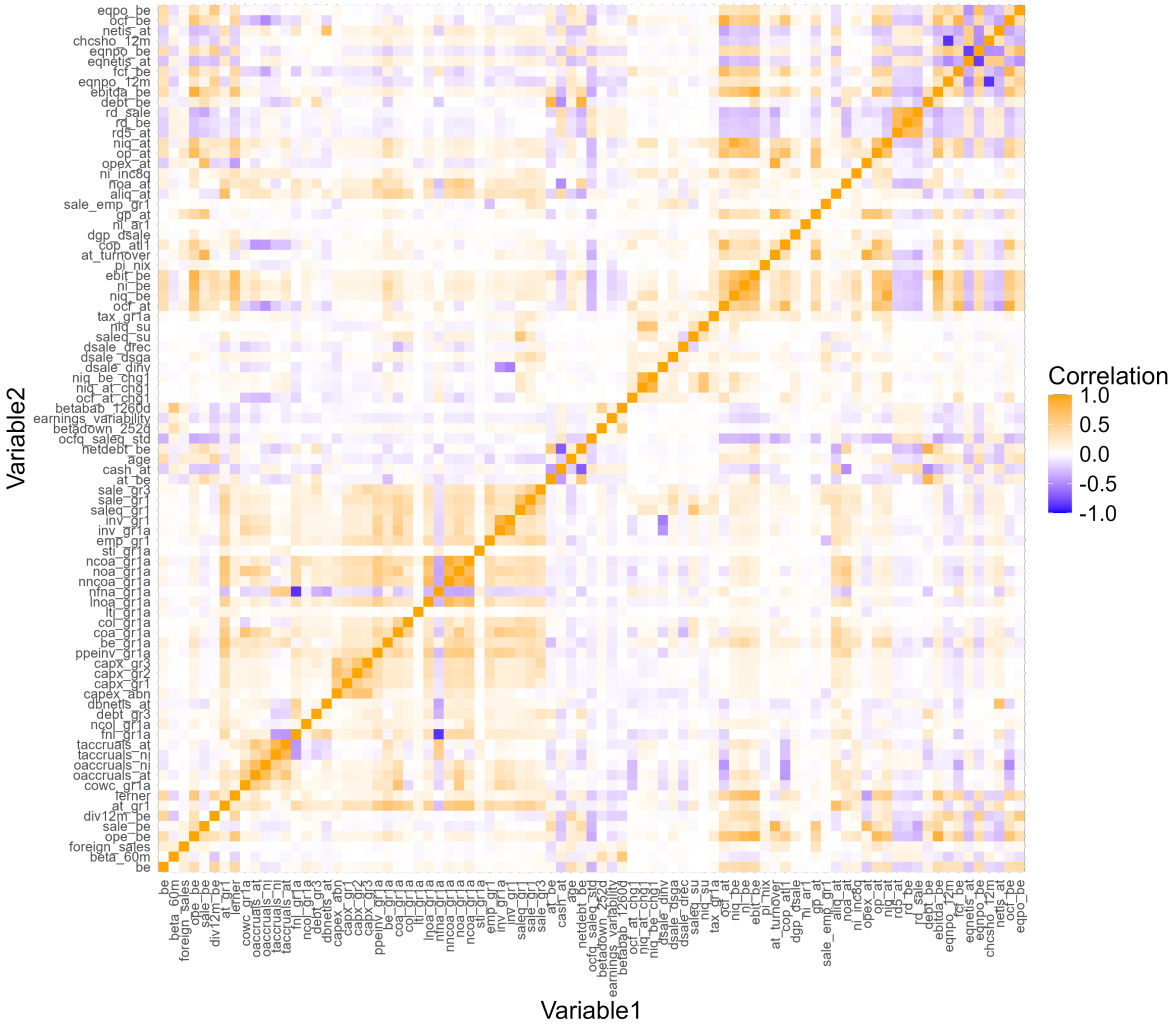


Figure A.2: Correlation matrix: Stock Characteristics

Note: Stock characteristics (except industry dummies) are rank-normalized cross-sectionally.

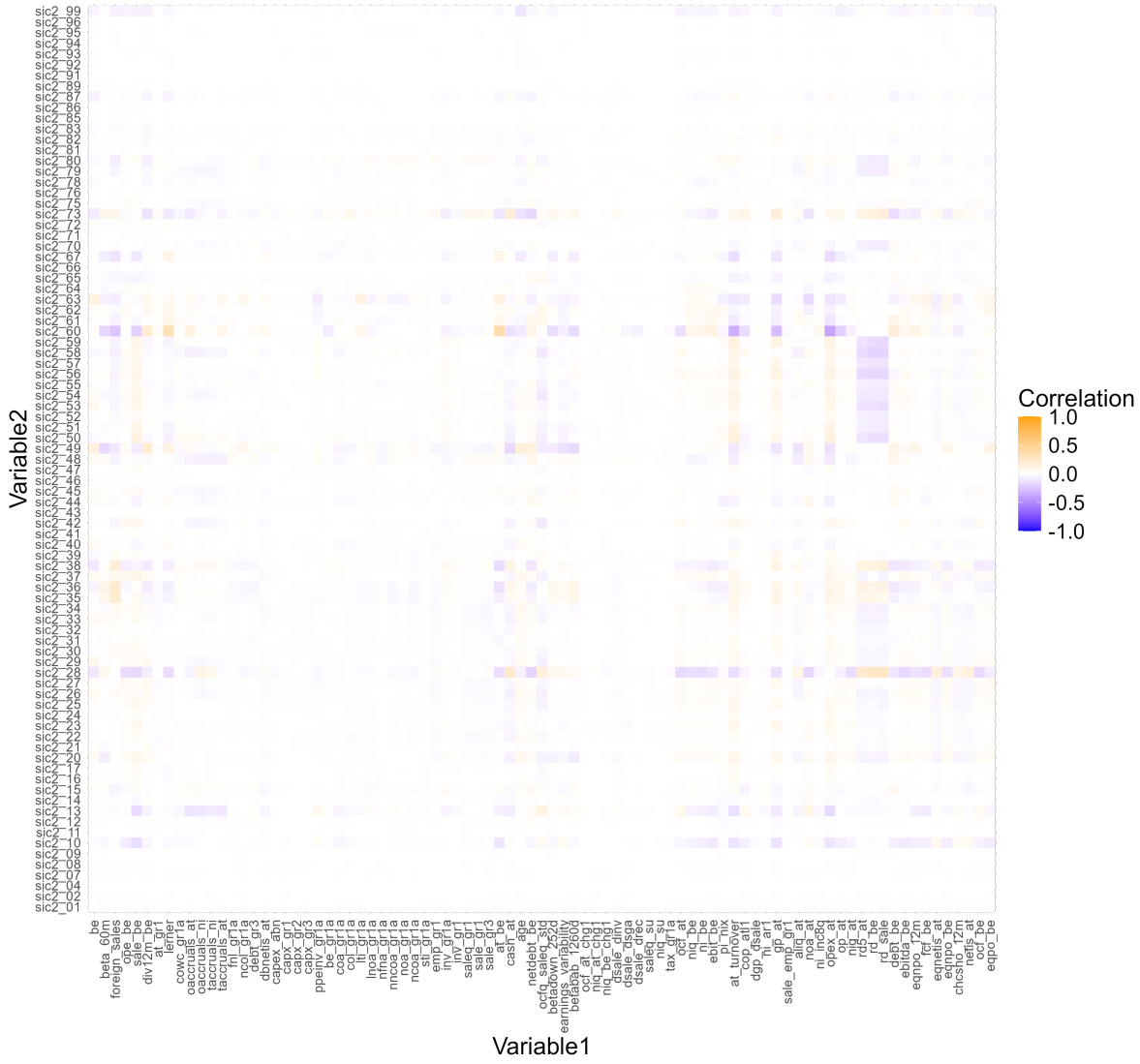


Figure A.3: Correlation matrix: Stock Characteristics vs Industries

Note: Stock characteristics (except industry dummies) are rank-normalized cross-sectionally.

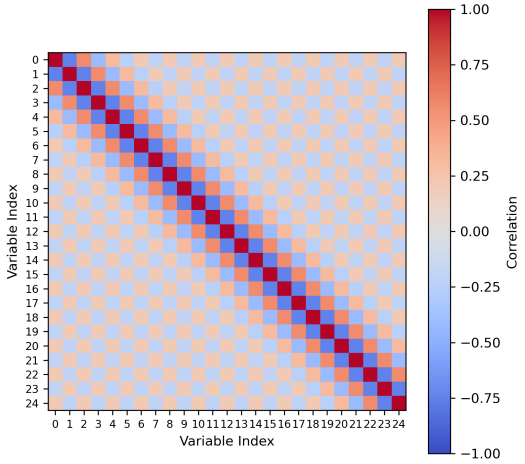


Figure B.1: Summary of data generating process in simulation study

Note:

B Details of Simulation Study Design

Assets are simulated. Their exogenous characteristics (other than market equity me) follow multivariate normal distribution:

$$x_i \sim N(0, \Sigma) \tag{33}$$

where $\Sigma \in \mathbb{R}^{p \times p}$ follows Toeplitz design with base $\sigma_{X,base} = -0.75$ truncated at $\underline{\sigma}_X = 0.2$. Specifically,

$$\sigma_{X,k,m} = \begin{cases} (-0.75)^{|k-m|} & \text{if } |(-0.75)^{|k-m|}| > 0.2 \\ 0.2 \cdot sign(k-m) & \text{otherwise} \end{cases} \tag{34}$$

The negative base $\sigma_{base} = -0.75$ induces negative (positive) correlation between characteristics whose indices differ by odd (even) number.[24] Since $(-0.75)^{|k-m|} = 1$ when $k = m$, the variance of asset characteristics is normalized to 1. As can be seen from (34), the key feature of Toeplitz-designed covariance matrix is that the (absolute) correlation between characteristics is decaying in the distance between the indices

[24]For example, if $k = 1$ and $m = 2$, one obtains $\sigma_{k,m} = (-0.75)^{|-1|} = -0.75 < 0$. However, for m that are distant by even number from k, the correlation is positive. For example, for $m = 3$, one has $\sigma_{k,m} = (-0.75)^{|-2|} = (-0.75)^2 > 0$.

of characteristics. Indeed, as $|k-m|$ becomes large, $(-0.75)^{|k-m|}$ goes to zero. To induce non-sparsity in Σ, I truncate the absolute correlation between any two asset characteristics at $\underline{\sigma} = 0.2$. This provides a more challenging setting for estimators since all characteristics are correlated at least to some degree. Toeplitz design of variance-covariance matrix is commonly used in simulation studies.[25] The illustration of Σ for $p = 25$ is provided in Figure B.1.

The market equity of asset me_j is generated to be endogenous:

$$me_j = \mu^T x_j + \eta z_j + \nu_j \tag{35}$$

where ν_j is endogenous component of market equity, whereas z_j is exogenous component. The correlation between market equity and exogenous characteristics is modelled via the vector $\mu \in \mathbb{R}^{p\times 1}$. The vector μ is drawn from p-dimensional uniform distribution $Unif(-0.5, 0.5)$ for each fund separately.[26] The parameter η governs the strength of the first stage[27] and is set to $\eta = 0.5$.

[25] See, for example Zou and Hastie (2005), Zou and Zhang (2009).

[26] This allows to account for the possibility that funds can have different investment universes, and for different investment universes, μ can be different.

[27] covariance between instrument z_j and endogenous variable me_j

C Proofs

C.1 Proof of Lemma 1.

Start with the sample counterpart of the moment conditions of the nonlinear GMM estimated on the *augmented* dataset:

$$\hat{\mathbb{E}}\left[z_{A,i,t,j}\left(\hat{\delta}_{A,i,t,j}\exp(-x_{A,i,t,j}^{T}\beta_{i,t})-1\right)\right]=0\in\mathbb{R}^{p\times 1} \tag{36}$$

Note that $Z_{A,i,t}=\begin{bmatrix}Z_{i,t}\\ \Psi_{i,t}\\ -\Psi_{i,t}\end{bmatrix}\in\mathbb{R}^{(n+2p)\times p}$, $X_{A,i,t}=\begin{bmatrix}X_{i,t}\\ \Psi_{i,t}\\ -\Psi_{i,t}\end{bmatrix}\in\mathbb{R}^{(n+2p)\times p}$,

and $\delta_{A,i,t}=\begin{bmatrix}\hat{\delta}_{i,t}\\ \exp\left(\Psi_{i,t}\beta_{i,t}^{target}\right)\\ \exp\left(-\Psi_{i,t}\beta_{i,t}^{target}\right)\end{bmatrix}\in\mathbb{R}^{(n+2p)\times 1}$, where n is the number of original (true) observations of holdings and p is the number of asset characteristics in the specification of the demand function. One can re-write the moment conditions (36) as:

$$\begin{aligned}
&\hat{\mathbb{E}}\left[z_{A,i,t,j}\left(\delta_{A,i,t,j}\exp(-x_{A,i,t,j}^{T}\beta_{i,t})-1\right)\right]\\
=&\frac{1}{n+2p}\sum_{j=1}^{n+2p}\left[z_{A,i,t,j}\left(\delta_{A,i,t,j}\exp(-x_{A,i,t,j}^{T}\beta_{i,t})-1\right)\right]\\
=&\frac{n}{n+2p}\cdot\frac{1}{n}\sum_{j=1}^{n}\left[z_{t,j}\left(\hat{\delta}_{i,t,j}\exp(-x_{t,j}^{T}\beta_{i,t})-1\right)\right]\\
+&\frac{1}{n+2p}\cdot\sum_{j=1}^{p}\left[\psi_{i,t,j}\left(\exp(\psi_{i,t,j}^{T}\beta_{i,t}^{target})\exp(-\psi_{i,t,j}^{T}\beta_{i,t})-1\right)\right]\\
+&\frac{1}{n+2p}\cdot\sum_{j=1}^{p}\left[-\psi_{i,t,j}\left(\exp(-\psi_{i,t,j}^{T}\beta_{i,t}^{target})\exp(\psi_{i,t,j}^{T}\beta_{i,t})-1\right)\right]\\
=&\frac{n}{n+2p}\cdot\frac{1}{n}\sum_{j=1}^{n}\left[z_{t,j}\left(\hat{\delta}_{i,t,j}\exp(-x_{t,j}^{T}\beta_{i,t})-1\right)\right]\\
+&\frac{1}{n+2p}\cdot\sum_{j=1}^{p}\left[\psi_{i,t,j}\left(\exp(\psi_{i,t,j}^{T}(\beta_{i,t}^{target}-\beta_{i,t}))-1\right)\right]\\
+&\frac{1}{n+2p}\cdot\sum_{j=1}^{p}\left[-\psi_{i,t,j}\left(\exp(-\psi_{i,t,j}^{T}(\beta_{i,t}^{target}-\beta_{i,t}))-1\right)\right]=0
\end{aligned} \tag{37}$$

Since $\Psi_{i,t} = \lambda_{i,t}^{synth} \cdot \mathrm{I}_p$, we have that $\psi_{i,t,j} = \lambda_{i,t}^{synth} \cdot \iota_j$, where $\iota_j \in \mathbb{R}^{p \times 1}$ is a vector such that elements $\iota_{j,k} = 1$ if $j = k$, and $\iota_{j,k} = 0 \ \forall j \neq k$. Then, the branch of the data augmentation-driven penalty arising from the positive-valued synthetic assets $\Psi_{i,t}$ can be simplified as:

$$\begin{aligned}
\sum_{j=1}^{p} \left[\psi_{i,t,j} \left(\exp(\psi_{i,t,j}^T(\beta_{i,t}^{target} - \beta_{i,t})) - 1\right)\right] =& \sum_{j=1}^{p} \lambda_{i,t}^{synth} \cdot \iota_j \left(\exp(\lambda_{i,t}^{synth} \cdot \iota_j^T(\beta_{i,t}^{target} - \beta_{i,t})) - 1\right) \\
=& \sum_{j=1}^{p} \lambda_{i,t}^{synth} \cdot \iota_j \left(\exp(\lambda_{i,t}^{synth}(\beta_{j,i,t}^{target} - \beta_{j,i,t})) - 1\right) \\
=& \begin{bmatrix} \lambda_{i,t}^{synth} \left(\exp(\lambda_{i,t}^{synth}(\beta_{1,i,t}^{target} - \beta_{1,i,t})) - 1\right) \\ \lambda_{i,t}^{synth} \left(\exp(\lambda_{i,t}^{synth}(\beta_{2,i,t}^{target} - \beta_{2,i,t})) - 1\right) \\ \cdots \\ \lambda_{i,t}^{synth} \left(\exp(\lambda_{i,t}^{synth}(\beta_{p,i,t}^{target} - \beta_{p,i,t})) - 1\right) \end{bmatrix} \\
=& \lambda_{i,t}^{synth} \left(\exp(\lambda_{i,t}^{synth}(\beta_{i,t}^{target} - \beta_{i,t})) - 1_p\right)
\end{aligned} \tag{38}$$

where in the last row, exp() is applied to the vector element-wise, and 1_p is a $p \times 1$ vector of ones. Analogously, for the negative-valued synthetic assets $-\Psi_{i,t}$, one obtains:

$$\begin{aligned}
\sum_{j=1}^{p} \left[-\psi_{i,t,j} \left(\exp(-\psi_{i,t,j}^T(\beta_{i,t}^{target} - \beta_{i,t})) - 1\right)\right] =& - \lambda_{i,t}^{synth} \left(\exp(-\lambda_{i,t}^{synth}(\beta_{i,t}^{target} - \beta_{i,t})) - 1_p\right) \\
=& - \lambda_{i,t}^{synth} \left(\exp(\lambda_{i,t}^{synth}(\beta_{i,t} - \beta_{i,t}^{target})) - 1_p\right)
\end{aligned} \tag{39}$$

Then, by (37), (38), and (39):

$$\begin{aligned}
& \hat{\mathbb{E}}\left[z_{A,i,t,j} \left(\delta_{A,i,t,j} \exp(-x_{A,i,t,j}^T \beta_{i,t}) - 1\right)\right] \\
=& \frac{n}{n+2p} \cdot \frac{1}{n} \sum_{j=1}^{n} \left[z_{t,j} \left(\hat{\delta}_{i,t,j} \exp(-x_{t,j}^T \beta_{i,t}) - 1\right)\right] \\
+& \frac{1}{n+2p} \cdot \lambda_{i,t}^{synth} \left(\exp(\lambda_{i,t}^{synth}(\beta_{i,t}^{target} - \beta_{i,t})) - 1_p\right) \\
-& \frac{1}{n+2p} \cdot \lambda_{i,t}^{synth} \left(\exp(\lambda_{i,t}^{synth}(\beta_{i,t} - \beta_{i,t}^{target})) - 1_p\right) = 0
\end{aligned}$$

By multiplying both sides by $\frac{n+2p}{n}$, we have:

$$\begin{aligned}
&\frac{1}{n}\sum_{j=1}^{n}\left[z_{t,j}\left(\hat{\delta}_{i,t,j}\exp(-x_{t,j}^{T}\beta_{i,t})-1\right)\right] \\
&\quad+\frac{1}{n}\cdot\lambda_{i,t}^{synth}\left(\exp(\lambda_{i,t}^{synth}(\beta_{i,t}^{target}-\beta_{i,t}))-1_p\right) \\
&\quad-\frac{1}{n}\cdot\lambda_{i,t}^{synth}\left(\exp(\lambda_{i,t}^{synth}(\beta_{i,t}-\beta_{i,t}^{target}))-1_p\right)=0
\end{aligned} \tag{40}$$

Note that in (40), the second term (third term) comes from the positive-valued (negative-valued) synthetic assets. Hence, the penalty induced by positive-valued synthetic assets can be expressed as:

$$\pi_{+}^{synth}\left(\lambda_{i,t}^{synth},\,\beta_{i,t}-\beta_{i,t}^{target}\right):=\frac{1}{n}\cdot\lambda_{i,t}^{synth}\left(\exp(\lambda_{i,t}^{synth}(\beta_{i,t}^{target}-\beta_{i,t}))-1_p\right)$$

while the penalty induced by negative-valued synthetic assets:

$$\pi_{-}^{synth}\left(\lambda_{i,t}^{synth},\,\beta_{i,t}-\beta_{i,t}^{target}\right):=-\frac{1}{n}\cdot\lambda_{i,t}^{synth}\left(\exp(\lambda_{i,t}^{synth}(\beta_{i,t}-\beta_{i,t}^{target}))-1_p\right)$$

Which can be re-written as:

$$\begin{aligned}
&\frac{1}{n}\sum_{j=1}^{n}\left[z_{t,j}\left(\hat{\delta}_{i,t,j}\exp(-x_{t,j}^{T}\beta_{i,t})-1\right)\right] \\
&\quad+\underbrace{\frac{1}{n}\cdot\lambda_{i,t}^{synth}\left(\exp(\lambda_{i,t}^{synth}(\beta_{i,t}^{target}-\beta_{i,t}))-\exp(\lambda_{i,t}^{synth}(\beta_{i,t}-\beta_{i,t}^{target}))\right)}_{:=\pi_{synth}\left(\lambda_{synth},\,\beta_{i,t}-\beta_{i,t}^{target}\right)}=0
\end{aligned} \tag{41}$$

Setting $\pi_{synth}\left(\lambda_{synth},\,\beta_{i,t}-\beta_{i,t}^{target}\right)$ to the second term in (41) finalizes the proof of Lemma 1. □

C.2 Proof of Lemma 2

Take from Lemma 1 the expressions for $\pi_{+}^{synth}\left(\lambda_{i,t}^{synth},\,\beta_{i,t}-\beta_{i,t}^{target}\right)$, $\pi_{-}^{synth}\left(\lambda_{i,t}^{synth},\,\beta_{i,t}-\beta_{i,t}^{target}\right)$, and $\pi_{synth}\left(\lambda_{synth},\,\beta_{i,t}-\beta_{i,t}^{target}\right)$.

Then, $\forall \lambda_{i,t}^{synth} > 0$, one obtains for positive-valued synthetic assets:

$$\lim_{(\beta_{k,i,t}-\beta_{k,i,t}^{target})\to+\infty} \left\| \frac{1}{n}\lambda_{i,t}^{synth} \left(\exp(-\lambda_{i,t}^{synth}(\beta_{k,i,t}-\beta_{k,i,t}^{target})) - 1 \right) \right\|_2 = \left\| -\frac{1}{n}\lambda_{i,t}^{synth} \right\|_2 = \frac{1}{n}\lambda_{i,t}^{synth}$$

$$\lim_{(\beta_{k,i,t}^{target}-\beta_{k,i,t})\to+\infty} \left\| \frac{1}{n}\lambda_{i,t}^{synth} \left(\exp(-\lambda_{i,t}^{synth}(\beta_{k,i,t}-\beta_{k,i,t}^{target})) - 1 \right) \right\|_2 = \|+\infty\|_2 = +\infty$$

Analogously, for the branch of penalty with negative-valued synthetic assets:

$$\lim_{(\beta_{k,i,t}-\beta_{k,i,t}^{target})\to+\infty} \left\| -\frac{1}{n}\lambda_{i,t}^{synth} \left(\exp(\lambda_{i,t}^{synth}(\beta_{k,i,t}-\beta_{k,i,t}^{target})) - 1 \right) \right\|_2 = \|-\infty\|_2 = +\infty$$

$$\lim_{(\beta_{k,i,t}^{target}-\beta_{k,i,t})\to+\infty} \left\| -\frac{1}{n}\lambda_{i,t}^{synth} \left(\exp(\lambda_{i,t}^{synth}(\beta_{k,i,t}-\beta_{k,i,t}^{target})) - 1 \right) \right\|_2 = \left\| \frac{1}{n}\lambda_{i,t}^{synth} \right\|_2 = \frac{1}{n}\lambda_{i,t}^{synth}$$

At the same time, the penalty based on both positive- and negative-valued synthetic assets exhibits:

$$\begin{aligned}
&\lim_{(\beta_{k,i,t}-\beta_{k,i,t}^{target})\to+\infty} \left\| \frac{1}{n}\lambda_{i,t}^{synth} \left(\exp(\lambda_{i,t}^{synth}(\beta_{k,i,t}^{target}-\beta_{k,i,t})) - \exp(\lambda_{i,t}^{synth}(\beta_{k,i,t}-\beta_{k,i,t}^{target})) \right) \right\|_2 \\
&\qquad = \left\| \frac{1}{n}\lambda_{i,t}^{synth} (0-\infty) \right\|_2 = +\infty \\
&\lim_{(\beta_{k,i,t}^{target}-\beta_{k,i,t})\to+\infty} \left\| \frac{1}{n}\lambda_{i,t}^{synth} \left(\exp(\lambda_{i,t}^{synth}(\beta_{k,i,t}^{target}-\beta_{k,i,t})) - \exp(\lambda_{i,t}^{synth}(\beta_{k,i,t}-\beta_{k,i,t}^{target})) \right) \right\|_2 \\
&\qquad = \left\| \frac{1}{n}\lambda_{i,t}^{synth} (\infty-0) \right\|_2 = +\infty
\end{aligned} \tag{42}$$

□

C.3 Proof of Proposition 1

The first part of proposition stating:

$$\hat{\mathbb{E}} \left[z_{A,i,t,j} \left(\hat{\delta}_{A,i,t,j} \exp(-x_{A,i,t,j}^T \beta_{i,t}) - 1 \right) \right] = 0$$

$$\Leftrightarrow$$

$$\frac{1}{n}\sum_{j=1}^{n} \left[z_{i,t,j} \left(\hat{\delta}_{i,t,j} \exp(-x_{i,t,j}^T \beta_{i,t}) - 1 \right) \right] + \pi_{synth} \left(\lambda_{i,t}^{synth},\, \beta_{i,t} - \beta_{i,t}^{target} \right) = 0$$

with:

$$\pi_{synth}\left(\lambda_{i,t}^{synth},\, \beta_{i,t} - \beta_{i,t}^{target}\right) = \frac{1}{n} \cdot \lambda_{i,t}^{synth}\left(\exp(\lambda_{i,t}^{synth}(\beta_{i,t}^{target} - \beta_{i,t})) - \exp(\lambda_{i,t}^{synth}(\beta_{i,t} - \beta_{i,t}^{target}))\right)$$

stems directly from (41) in Lemma 1.

In the second part of proposition, there are 4 properties. I show them one-by-one below.

1. From Lemma 1, if $\lambda_{i,t}^{synth} = 0$, then:

$$\pi_{synth}\left(\lambda_{i,t}^{synth},\, \beta_{i,t} - \beta_{i,t}^{target}\right) = \frac{1}{n} \cdot 0 \cdot (1_p - 1_p) = 0 \in \mathbb{R}^{p\times 1}$$

where 1_p denotes $p \times 1$ vector of ones.

2. Analogously, if for a given $k \in \{1, ..., p\}$, $\beta_{k,i,t} = \beta_{k,i,t}^{target}$, then one obtains that penalty on k-th coefficient:

$$\pi_{synth}\left(\lambda_{i,t}^{synth},\, \beta_{k,i,t} - \beta_{k,i,t}^{target}\right) = \frac{1}{n} \cdot \lambda_{i,t}^{synth} \cdot (1 - 1) = 0 \in \mathbb{R}^{1\times 1}$$

3. Note that if $\beta_{k,i,t} \neq \beta_{k,i,t}^{target}$, then either $\beta_{k,i,t} > \beta_{k,i,t}^{target}$ or $\beta_{k,i,t} < \beta_{k,i,t}^{target}$. If $\beta_{k,i,t} > \beta_{k,i,t}^{target}$, then as $\lambda_{i,t}^{synth} \to +\infty$:

$$\left(\exp(\lambda_{i,t}^{synth}(\beta_{k,i,t}^{target} - \beta_{k,i,t})) - \exp(\lambda_{i,t}^{synth}(\beta_{k,i,t} - \beta_{k,i,t}^{target}))\right) \to -\infty$$

Analogously, if $\beta_{k,i,t} < \beta_{k,i,t}^{target}$, then as $\lambda_{i,t}^{synth} \to +\infty$:

$$\left(\exp(\lambda_{i,t}^{synth}(\beta_{k,i,t}^{target} - \beta_{k,i,t})) - \exp(\lambda_{i,t}^{synth}(\beta_{k,i,t} - \beta_{k,i,t}^{target}))\right) \to +\infty$$

To complete the proof of the property, note that in both cases the norm will go to infinity.

4. Note that $|\beta_{k,i,t} - \beta_{k,i,t}^{target}| \to \infty$ can be decomposed into two cases:

1) $\beta_{k,i,t} - \beta_{k,i,t}^{target} \to +\infty$

2) $\beta_{k,i,t}^{target} - \beta_{k,i,t} \to +\infty$

Then, the last property in Proposition 1 stems directly from (42) in Lemma 2.

□

D Estimated Price Elasticities

Variable	Nonlinear GMM (Group, LD)	Pre-data augm. (Group, LD)	Pre-data augm. (Group, HD)
Nonlinear GMM (Group, LD)	1.000	0.734	0.979
Pre-data augm. (Group, LD)	0.734	1.000	0.726
Pre-data augm. (Group, HD)	0.979	0.726	1.000

Table D.1: Correlation between estimated coefficients on market equity

Note: This table reports the correlation between the estimates of demand function coefficient on market equity $\hat{\theta}_{i,t}$ in (1) obtained using the following three methods: 1) Nonlinear GMM of Koijen et al. (2023) estimated at the group-level with low-dimensional set of stock characteristics; 2) pre-data augmentation Debiased GMM under orthogonal moment conditions of Chernozhukov et al. (2018) estimated at the group-level with low-dimensional set of stock characteristics; 3) pre-data augmentation Debiased GMM estimated at the group-level with high-dimensional set of stock characteristics. For each of the three methods, the groups are formed following the algorithm of Koijen et al. (2023): investors of the same type (here, only one type: active mutual funds) are ranked by AUM and grouped so that each group has at least 2000 observations (including zero-weight observations). For all three methods, the estimation is performed at the quarterly frequency. The sample constitutes of active equity mutual funds in the U.S. and spans from 1990 Q1 to 2022 Q4. Price elasticity can be computed as approximately $1 - \hat{\theta}_{i,t}$ (Koijen et al. (2023)).

Figure D.1 shows the time-series evolution of the cross-sectional averages (Panel A) as well as percentiles of the cross-sectional distribution (Panel B) of the estimates for the coefficient on market equity $\hat{\theta}_{i,t}$ from (1). Given that the price elasticity of investor i's demand at time t can be approximately computed as $1 - \hat{\theta}_{i,t}$, it easy to interpret the magnitude of $\hat{\theta}_{i,t}$. Consistent with findings in Koijen et al. (2023) for small active investment advisors,[28] the AUM-weighted cross-sectional averages of price elasticity are in ballpark of 0.5. Cross-sectional distribution is also in a similar range to Koijen et al. (2023), with 90th percentile having close to unit elasticity and 10th percentile having price elasticity of around 0.3-0.4.[29]

Notably, all three methods produce similar estimates of price elasticity: The correlation between the estimates obtained using nonlinear GMM of Koijen et al. (2023) and the Debiased GMM under the (immunized to high-dimensionality) orthogonal moment conditions of Chernozhukov et al. (2018) is 0.73 (see Table D.1). The very high correlation of 0.98 between the low-dimensional and high-dimensional Debiased GMM estimates for the price elasticity suggests that controlling for the extended set of stock characteristics does not have first order importance the estimates of price elasticity.

[28] Koijen et al. (2023) estimate demand system on 13F holdings data and do not separate mutual funds from the overall group of investment advisors. Given that long-term investors (such as pension funds and insurance companies) and hedge funds are excluded from the definition of investment advisors in Koijen et al. (2023), active investment advisors can be viewed as a group of investors that largely contains and is similar to active mutual funds in my sample.

[29] Note that since price elasticity is approximately $1 - \hat{\theta}_{i,t}$, the 10th percentile in terms of price elasticity translates into 90th percentile in terms of estimated $\hat{\theta}_{i,t}$.

One can note that there is no major patterns in the evolution of the price elasticity for active equity mutual funds over time, except for the latter part of the period (after 2015), where AUM-weighted average price elasticity remains consistently slightly below 0.5. Notably, the equally-weighted average and median price elasticities are higher than the AUM-weighted average, suggesting that larger mutual funds tend to have lower price elasticity. This is consistent with the economic rationale that larger mutual funds pursue more passive investment strategies due to their larger diseconomies of scale.

D.1 First stage

To evaluate the relevance condition of the instrument for market equity me^{IV} in (4), I estimate the first stage where the (endogenous) market equity is regressed on the proposed by Koijen and Yogo (2019), Koijen et al. (2023) instrument and other (exogenous) stock characteristics:

$$me_{j,t} = \kappa_{g,t} me^{IV}_{i,t,j} + \zeta^T_{g,t} x_{j,t} + \epsilon_{i,t,j} \ , \ i \in g \tag{43}$$

where $me^{IV}_{i,t,j}$ is defined in (4). Notably, both the instrument and the first stage are investor-specific for two reasons. First, the instrument $me^{IV}_{i,t,j}$ is based on the investment universes of all institutional investors, except investor i. Second, the set of stocks j that comprises investor i's investment universe in different for each investor i. For the benchmark method of Koijen et al. (2023), I estimate the first stage (43) using OLS without penalty on coefficients. To ensure the precision of estimation of the first stage, I follow Koijen et al. (2023) and group mutual funds into groups $g \in \mathcal{G}$ so that each group has at least 2000 observations (including zero-weight observations). For the pre-data augmentation IV estimation, I employ Debiased GMM under orthogonal moment conditions of Chernozhukov et al. (2018) estimated at the group-level. Note that it is only the IV estimation that is always performed on 2000-observation grouped sample: the demand function loadings $\beta_{i,t}$ are estimated fully . Since Debiased GMM is robust to high-dimensionality and overfitting, I apply it to estimation of the first stage for both low-dimensional and high-dimensional specifications of (43). Figure D.2 plots the distribution of the mutual fund-quarter t-statistics on the coefficient $\kappa_{i,t}$ in

(43).[30] All three methods produce strong first stage, with practically all of the t-statistics being above the critical value of 4.05 for rejecting the null of weak instruments at 5% significance level (Stock and Yogo (2005)). These results support the relevance of the instrument for market equity defined in (4) for estimation of demand function specification in (1).

[30]Individual fund estimates $\kappa_{i,t}$ are equal to the group level estimates $\kappa_{g,t}$ to which fund i belongs.

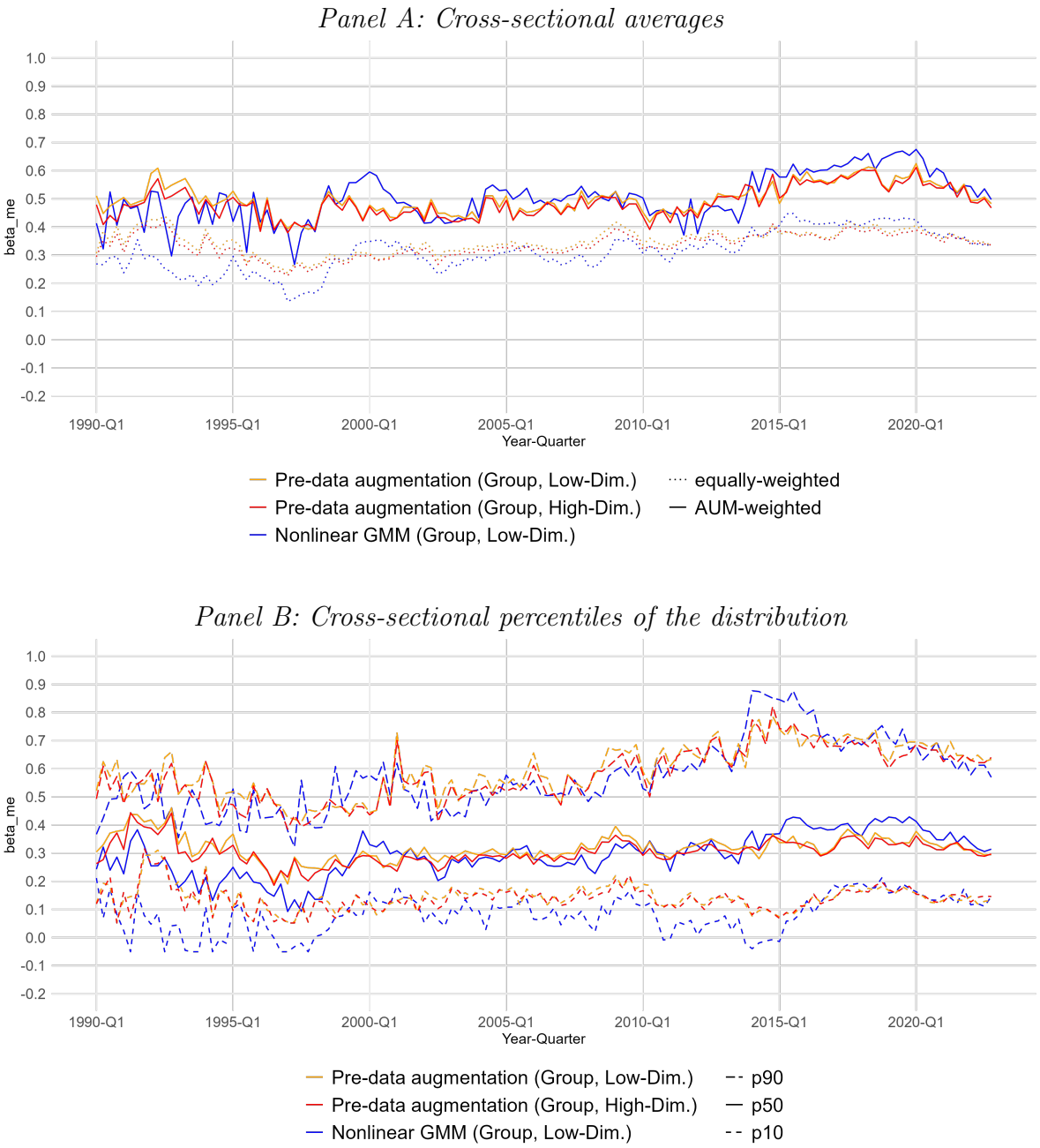


Figure D.1: Time-series evolution of the estimated coefficients on market equity

Note: This figure reports the time-series evolution of the estimates of demand function coefficient on market equity $\hat{\theta}_{i,t}$ in (1) obtained using the following three methods: 1) Nonlinear GMM of Koijen et al. (2023) estimated at the group-level with low-dimensional set of stock characteristics; 2) pre-data augmentation Debiased GMM under orthogonal moment conditions of Chernozhukov et al. (2018) estimated at the group-level with low-dimensional set of stock characteristics; 3) pre-data augmentation Debiased GMM estimated at the group-level with high-dimensional set of stock characteristics. For each of the three methods, the groups are formed following the algorithm of Koijen et al. (2023): investors of the same type (here, only one type: active mutual funds) are ranked by AUM and grouped so that each group has at least 2000 observations (including zero-weight observations). For all three methods, the estimation is performed at the quarterly frequency. The sample constitutes of active equity mutual funds in the U.S. and spans from 1990 Q1 to 2022 Q4. Price elasticity can be computed as approximately $1 - \hat{\theta}_{i,t}$ (Koijen et al. (2023)).

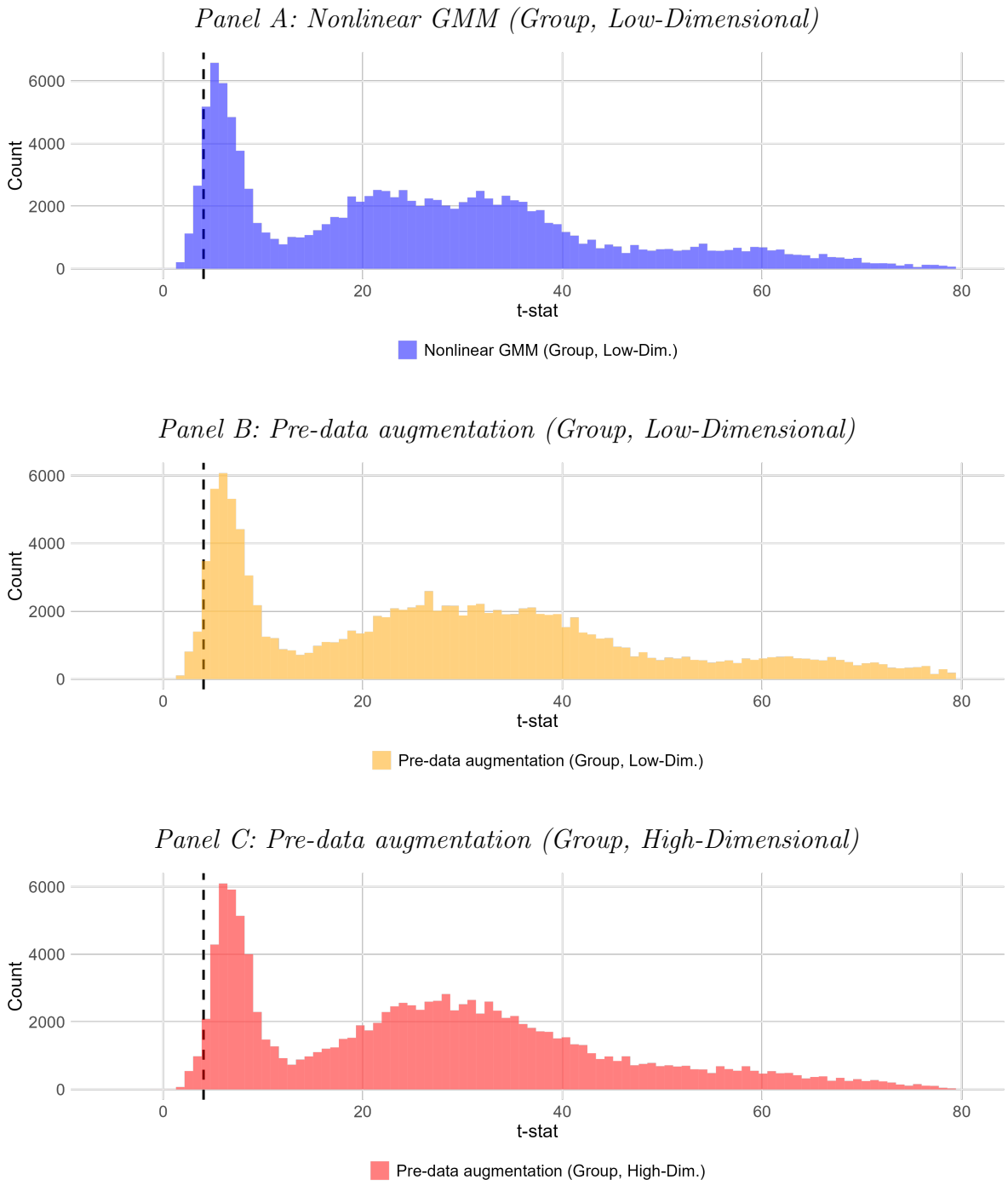


Figure D.2: IV for market equity: first stage

Note: The figure shows the histogram of t-statistics from the first stage of IV estimation of investors' demand functions in (1) for each of the following three methods: 1) Nonlinear GMM of Koijen et al. (2023) estimated at the group-level with low-dimensional set of stock characteristics; 2) pre-data augmentation Debiased GMM under orthogonal moment conditions of Chernozhukov et al. (2018) estimated at the group-level with low-dimensional set of stock characteristics; 3) pre-data augmentation Debiased GMM estimated at the group-level with high-dimensional set of stock characteristics. For each of the three methods, the groups are formed following the algorithm of Koijen et al. (2023): investors of the same type (here, only one type: active mutual funds) are ranked by AUM and grouped so that each group has at least 2000 observations (including zero-weight observations). The vertical black dashed line denotes the critical value of 4.05 for rejecting the null of weak instruments at 5% significance level (Stock and Yogo (2005)). For all three methods, the estimation is performed at the quarterly frequency. The sample constitutes of active equity mutual funds in the U.S. and spans from 1990 Q1 to 2022 Q4.