


# Uncovering Sparsity and Heterogeneity in Firm-Level Return Predictability Using Machine Learning

Theodoros Evgeniou  
*INSEAD Decision Sciences*  
theodoros.evgeniou@insead.edu

Ahmed Guecioueur   
*INSEAD Finance*  
ahmed.guecioueur@insead.edu (corresponding author)

Rodolfo Prieto  
*INSEAD Finance*  
rodolfo.prieto@insead.edu

## Abstract

We develop an approach that combines the estimation of monthly firm-level expected returns with an assignment of firms to (possibly) latent groups, both based on observable characteristics, using machine learning principles with linear models. The best-performing methods are flexible two-stage sparse models that capture group-membership predictive relationships. Portfolios formed to exploit such group-varying predictions based on a parsimonious set of characteristics deliver economically meaningful returns with low turnover. We propose statistical tests based on nonparametric bootstrapping for our results, and detail how different characteristics may matter for different groups of firms, making comparisons to the existing literature.

## I. Introduction

New methods for predictive modeling of data sets with large cross-sectional or time-series dimensions (or both) have in recent years been developed in the field of machine learning (ML). Stock return predictability is a natural target as it is a low signal-to-noise ratio problem with a high dimensional set of predictive signals. Specifically, ML methods can help in the search for the combination of conditioning variables and cross-sectional factors that best describes the returns of individual

---

We thank Jennifer Conrad (the editor) and Alberto Martín-Utrera (the referee) for their constructive comments. We are grateful to Panos Mavrokonstantis for excellent research assistance while he was a Senior Research Scientist at INSEAD. We also thank participants at the 13th Annual SoFiE Conference, the 3rd Future of Financial Information Conference, the inaugural Miami Herbert Winter Research Conference on ML and Business, the 2021 AFA PhD Poster Session, the 2020 European Winter Meetings of the Econometric Society, the 22nd INFER Annual Conference, the 9th Wharton-INSEAD Doctoral Consortium, and the INSEAD Accounting and Finance PhD seminar series, as well as Alex Chinco (discussant), Victor DeMiguel, Scott Murray (discussant), Joël Peress, Marcel Rindisbacher, Raman Uppal, Jinyuan Zhang, and Guofu Zhou for their helpful comments. A previous version of this article was circulated under the title “Modeling Heterogeneity in Firm-Level Return Predictability with Machine Learning.”

assets, freeing researchers from having to impose ad hoc sparsity on asset-pricing models by manually selecting a subset of factors or variables (Nagel (2021)).

Gu, Kelly, and Xiu (GKX) (2020) make an important contribution by introducing a wide variety of ML techniques (each with strengths and weaknesses) to measure the conditional mean function of firm-level stock excess returns. Their data set is large (a 60-year period of 30,000 stocks and over 900 predictors), and they find that ML methods provide substantial improvements in out-of-sample (OOS) predictability over OLS. Their best-performing model is a neural network with a small number of layers. However, despite the promising quantitative performance, economic interpretability remains a challenge for ML approaches and may be particularly severe for the more black-box-like nonlinear methods such as neural networks. As Karolyi and Van Nieuwerburgh (2020) emphasize, only with solid economic intuition can such exercises serve as the basis for more realistic asset-pricing theories.<sup>1</sup>

In this study, we take on the challenge laid down by Karolyi and Van Nieuwerburgh (2020) and adapt ML models to study heterogeneity in firm-level return predictability. Specifically, our objective is to tease out how group membership may determine firm-level risk premia.<sup>2</sup> The distinction between variables that define firm groupings and variables that predict returns will be one of the key insights of our approach. We are motivated by the fact that standard formulations assume away the possibility that asset returns are priced by risk factors that depend on a firm's group membership (Patton and Weller (2022)), a regularity that may be connected to the fact that different firm characteristics influence investors differently, signaling, for example, preferred risk habitats (Dorn and Huberman (2010), Balasubramaniam, Campbell, Ramadorai, and Ranish (2023)). We diverge from the approach taken by existing studies of firm-level return predictability that fit a single model to all firms (we label them "pooled" models) by estimating "by-group" models.<sup>3</sup>

Our design is purposely simple. We contrast two broad grouping criteria: First, firms are grouped according to their industry classifications, measured by SIC codes, and second, group membership is inferred using an unsupervised learning technique, *k*-means clustering.<sup>4</sup> All our models use firm characteristics and market-wide variables as predictors and are evaluated OOS as in GKX. We differ by using linear models as building blocks to construct economically motivated functional forms,<sup>5</sup> without needing to call upon the unrestricted flexibility of neural networks, for example. The role of ML in our study is that of a tool: for estimating our

<sup>1</sup>One route is to embed ML methods in equilibrium models, as Fuster, Goldsmith-Pinkham, Ramadorai, and Walther (2022) do; another route is to incorporate economic structure into the ML methods themselves, as we do.

<sup>2</sup>We use the terms "expected return" and "risk premium" interchangeably.

<sup>3</sup>Prior studies (Gu et al. (2020), Freyberger et al. (2020), and DeMiguel et al. (2020)) use various approaches assuming that the same set of variables matters for all firms in the cross section.

<sup>4</sup>Clustering groups similar firms together using a standard distance metric. We find that the number and interpretation of discernible clusters is stable.

<sup>5</sup>The linear stochastic discount factors adopted in the segmented market model of Patton and Weller (2022) are a variation of the linear factor models derived from Merton (1973) (intertemporal capital asset pricing model) and Ross (1976) (arbitrage pricing theory). Lustig, Van Nieuwerburgh, and Verdelhan (2013) and more recently Farmer, Schmidt, and Timmermann (2019) show how linear representations of expected excess returns arise in equilibrium.

predictive models in a high-dimensional setting, and for estimating a partition of the cross section. Our models are also transparently interpretable without the need for further assumptions or approximations: For example, we can directly interpret the estimated coefficient values of our Lasso-regularized models in terms of predictive variable importance. Furthermore, we test the relative predictive performance of different methods (and parameter estimates) by applying a bootstrap method to our panel of firms.

Our first finding is that the incremental OOS predictability of “by-group” over “pooled” models with the same regularization scheme is positive and statistically significant across the board. That is, identifying and employing an economic partition of the cross section can positively impact OOS firm-level predictive performance.

Building on this evidence, we develop flexible two-stage models that capture group-membership predictive relationships in addition to a common set of predictive relationships. A Ridge-regularized model for industry membership and a Lasso-regularized model for clusters are the best-performing models when assessed on the full cross section of firms. We detect substantial levels of predictability, with an OOS  $R^2$  metric of over 1% for the full cross section of firms, and over 1.9% when training on a subset of the largest firms. Although our sample of firms and time period studied are not the same as for GKX, our OOS  $R^2$ s are higher than their best-attained level of 0.4%.

Importantly, we find that sparse models consistently deliver better OOS predictive performance when clustering is used instead of industry membership to define firm groupings. The outperformance is highest for the by-group Lasso model and remains when both pooled and by-group stages are combined in our two-stage Lasso-regularized model; this is not necessarily the case for Ridge-regularized models. Coupled with the high overall OOS predictive performance of sparse models,<sup>6</sup> this evidence suggests that clustering the cross section to define heterogeneous predictive relationships uncovers sparsity in predictive variable importance while simultaneously delivering improved OOS predictive performance. This finding is consistent with the recent search for parsimony in the asset-pricing factors literature (Feng, Giglio, and Xiu (2020)), a parsimony that is in stark contrast to the existing literature on return predictability.

We confirm that our firm-level OOS predictability results translate into economically meaningful returns for portfolios. An investor forming portfolios based on the sign of next-month excess returns predicted by models that combine characteristics-based predictive heterogeneity with only a sparse set of selected predictive characteristics earns an annualized OOS Sharpe ratio (SR) of 1.18 (value-weighted (VW)) or 1.21 (equal-weighted (EW)). Portfolios formed based on models that ignore predictive heterogeneity always underperform compared with those that incorporate predictive heterogeneity. In our framework, the

<sup>6</sup>Sparsity has proved useful for empirical causal inference (Belloni et al. (2012), (2014), Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, and Newey (2017)) and theoretical modeling of economic behavior (Gabaix (2014), Hanna, Mullainathan, and Schwartzstein (2014), Gabaix (2020), and Guecioueur (2020)), where it can be related to broader forms of inattention (Sims (2003), Reis (2006), Chetty, Looney, and Kroft (2009), and Gabaix (2019)).

predicted returns are themselves a linear combination of the (potentially sparse) predictive characteristics selected by our models, without suffering from the multisignal selection issues highlighted by Novy-Marx (2016).

We propose a novel channel to answer Cochrane's (2011) questions of "Which characteristics really provide independent information about average returns? Which are subsumed by others?": Characteristics matter not only as predictors of next-period returns, but also in proxying for firms' latent group memberships. Furthermore, different and sparse sets of characteristics matter for different groups of firms. To illustrate how the distinction between grouping variables and predictive variables is one of the key insights of our approach, consider the AGE firm-level characteristic. Jiang, Lee, and Zhang (2005) found empirical evidence that the AGE variable predicts returns, yet this is unsupported by our findings. Rather, we find that for the cluster containing more mature firms, a few firm-level characteristics and market-level variables tend to be selected. Therefore, AGE is an important firm-level characteristic insofar as it helps us to identify a grouping of mature firms, rather than a predictor of firm-level returns itself. It is also notable that Balasubramaniam et al. (2023) found that AGE "has the strongest clientele effect" among (Indian) investors, suggesting a possible role that clientele effects may play in determining conditional risk premia.

Categories of grouping variables and predictive variables can help rationalize the apparent existence of many variables that appear to predict firm-level returns in previous work, but whose importance is of a different nature altogether. For example, our Lasso-regularized two-stage model selects only 7 firm-level characteristics and 1 market-level variable in total, and discards the rest as having no predictive worth. Low-frequency cash and profitability variables matter the most. In contrast, GKX (see also Freyberger et al. (2020)) find a large set of informative stock-level predictors that are notably absent from our Lasso-selected set: price trend variables (e.g., stock and industry momentum) followed by liquidity variables (e.g., market value and bid-ask spread) and volatility measures. Consistent with the selection of slow-moving fundamental variables as signals, portfolios formed using our methodology enjoy the benefit of a lower turnover compared with those tested by GKX.

## A. Related Literature

There is an extensive literature on forecasting *aggregate* market returns, typically measured by index returns. A notable example is Campbell and Thompson (2008). More recently, Rapach, Strauss, and Zhou (2010), Diebold and Shin (2019), and Rapach and Zhou (2020) have used ML techniques to produce forecast combinations for time series such as market returns or macroeconomic variables. Rapach et al. (2019) use similar techniques to forecast industry-level portfolio returns. Aggregating firms eliminates group-level heterogeneity; we have therefore not attempted to make predictions for market-level returns. Note, though, that OOS predictive accuracy is higher for market-level return prediction problems in the literature, so the problem we set out to tackle is the hardest-to-forecast setting.

The literature on forecasting *firm-level* returns is more recent and more relevant to our work. The closest empirical setup to ours is that of GKX. Han, He,

Rapach, and Zhou (2021) take a pure forecast combination approach that builds on the work of Lewellen (2015): The essence of such a forecast combination approach is to fit a set of models to either common or model-specific predictive variables, then have each model produce forecasts for all firms' next-period returns, and then finally combine these multiple forecasts of the same outcomes together (e.g., by the simple average). Our approach to heterogeneity is loosely related in that we link individual models together, but is distinct because we do not produce multiple forecasts for the same inputs: Our approach is not based on forecast combinations. Our focus is instead on detecting predictive relationships that are specific to groups of firms, so the individual models we link together vary in which subset of firms in the cross section they are applied to. Other approaches to predicting firm-level returns that do not involve forecast combinations are by Freyberger et al. (2020) and Fisher, Puelz, and Carvalho (2020), which both use spline-based regressions but differ in what procedures they use to select characteristics: The former uses the Group Lasso, whereas the latter takes a decision-theoretic approach. The parsimony that we uncover in characteristic importance has parallels with the parsimony uncovered among asset-pricing factors by Feng et al. (2020); we also exploit Lasso regularization for that (shared) goal. We share some objectives with DeMiguel, Martin-Utrera, Nogales, and Uppal (2020), who take a portfolio choice approach in order to incorporate the effects of transaction costs when determining characteristic importance.

Green, Hand, and Zhang (GHZ) (2017) attempt to discern which firm characteristics provide independent information about monthly stock returns by employing Fama–MacBeth-style contemporaneous regressions. Similarly, Kelly, Pruitt, and Su (2019) develop an intercept test that discriminates whether a characteristic-based return phenomenon is consistent with a beta/expected return model, using a method labeled instrumental principal components analysis (IPCA) that treats characteristics as instrumental variables for estimating dynamic loadings on latent factors.<sup>7</sup> Like these studies, we wish to understand the determinants of firm-level excess returns, and we do so by incorporating an economically grounded notion of heterogeneity in the relationships between characteristics and future returns that is amenable to testing.

Our study is related to recent work by Patton and Weller (2022), who study asset-level heterogeneity and segmentation, building upon the clustering techniques of Bonhomme and Manresa (2015); they do not focus on predictability, as we do. There have also been other applications of cluster analysis to the study of asset-pricing (Grishchenko and Rossi (2012), Ando and Bai (2017)) as well as more general unsupervised learning techniques (Gu, Kelly, and Xiu (2021)). Prior studies of heterogeneity in asset pricing consider contemporaneous relationships between firm characteristics and returns; we are not aware of any prior study that considers heterogeneity in predictive relationships, as we do.

<sup>7</sup>The methodology of Kelly et al. (2019) is also indirectly related to our own: Ding and He (2004) showed that principal components are the “continuous solutions” to the discrete group memberships that we assign to firms by the  $k$ -means clustering algorithm. We then differ in using characteristics to predict next-period firm-level returns given such a partition, whereas Kelly et al. (2019) use characteristics to infer contemporaneous factor loadings.

Another stream of asset-pricing literature that is important for our study relates to industry membership. Hou and Robinson (2006) found a contemporaneous relationship in the cross section between firms' industry memberships and financial returns. Daniel, Mota, Rottke, and Santos (2020) show how an optimal combination of characteristics-based factors and hedge portfolios delivers more efficient factors, and propose industry membership as one possible candidate of unpriced common variation. Our research using industry membership relates to Cohen and Frazzini (2008) and Menzly and Ozbas (2010), who find that economic links among certain individual firms and industries contribute significantly to cross-firm and cross-industry return predictability. Barrot and Sauvagnat (2016) examine whether firm-level idiosyncratic shocks propagate in production networks and add to a growing body of work in financial economics that studies how firms are affected by their customers and suppliers. Finally, a number of papers have found that effects that were studied in the overall market also exist when conditioning on industry membership, including Lewellen (1999), Moskowitz and Grinblatt (1999), Asness, Porter, and Stevens (2000), and Hou (2007).

The remainder of the article is organized as follows: We present our theoretical background in Section II and our empirical framework in Section III. We explain our data construction procedure in Section IV. In Section V, we apply our methodology to the data using different grouping specifications, and Section VI treats variable importance and heterogeneity. We conduct our portfolio analyses in Section VII. Section VIII concludes. The Supplementary Material provides a review of methods used in the article, as well as further data details and robustness checks.

## II. Background

We briefly introduce the theoretical motivation and ML methods we rely on to estimate our models.

### A. Heterogeneity and Predictability

Take a firm's next-month excess returns  $r_{i,t+1}$ , defined as the return in excess of the risk-free rate, as an additive prediction error model,

$$(1) \quad r_{i,t+1} = E_t(r_{i,t+1}) + \varepsilon_{i,t+1} = h(\mathbf{c}_{it}) + \varepsilon_{i,t+1},$$

where stocks are indexed as  $i = 1, \dots, N_t$ , months by  $t = 1, \dots, T$ , and  $\mathbf{c}_{it}$  is an  $M$ -dimensional vector of predictors, with an entry in  $\mathbf{c}_{it}$  corresponding to either a firm *characteristic* or a marketwide variable. Both types of state variables have been used extensively and will be described in detail in our empirical implementation. A linear model imposes that the conditional expectation  $E_t(r_{i,t+1})$  can be approximated by a linear function of the predictors,

$$(2) \quad h(\mathbf{c}_{it}) = \boldsymbol{\theta}' \mathbf{c}_{it},$$

where  $\boldsymbol{\theta} \in \mathbb{R}^M$  is a constant vector. Equation (2) can be traced back to the standard conditional CAPM pricing representation of expected returns,  $E_t(r_{i,t+1}) = \beta_{i,t}' f_t$ ,

which combines loadings  $\beta_{i,t}$  with marketwide factors  $f_t$ . These can be modeled explicitly as functions of an asset's own characteristics, as in Kelly et al. (2019) and Kojien and Yogo (2019).<sup>8</sup>

The standard conditional CAPM assumes away investor specialization and market segmentation, yet these are known to affect predictability. For example, Menzly and Ozbas (2010) show that information diffuses gradually in financial markets, impacting return predictability along vertical customer–supplier relationships across industries. More recently, Patton and Weller (2022) find evidence of segmentation across their choice of test assets, benchmark factor models, and time periods by studying heterogeneous market prices of risk using a CAPM of the form  $E_t(r_{i,t+1}) = \alpha_i + \beta_i'(f_t + \Phi_i \mathbf{I}_i)$ , where  $\mathbf{I}_i$  denotes a  $K$ -dimensional vector of indicator variables defining firm  $i$ 's group membership, with  $j \in \{1, \dots, K\}$ , and the matrix  $\Phi_i$  holds the deviation of risk premia for each group  $j$  from the vector of common factors  $f_t$ .

Using the latter as our backdrop, we set out to implement a representation of expected excess returns as a function of predictor variables using a class of linear models that exploits group heterogeneity. In particular, we augment equation (2) so that each firm  $i$  also belongs to a single group  $j$  (specified by a group membership variable  $z_i$  used by the indicator function  $\mathbb{1}\{z_i = j\}$ ) and with additional group-specific coefficients  $\theta_j$ ,

$$(3) \quad h_j(\mathbf{c}_{it}, z_i) = (\alpha_0 + \theta'_0 \mathbf{c}_{it}) + \sum_{j=1}^K (\alpha_j + \theta'_j \mathbf{c}_{it}) \mathbb{1}\{z_i = j\}.$$

In Section III.A, we explain how we estimate a version of this model.

## B. Machine Learning

With a high-dimensional characteristics vector  $\mathbf{c}_{it}$ , GKX found that ML methods outperformed traditional ones for the task of cross-sectional return predictability. We will use ML methods both to estimate predictive models and to infer firm groupings.

Our augmented linear equation (3) has multiple distinct vectors of coefficients  $\theta_0$  and  $\theta_j$ , and these are estimated based on different samples of firms (as we will explain in Section III). We employ ML estimation procedures that penalize/regularize (functions of) the norms of the coefficient vectors  $\theta_0$  and  $\theta_j$ . The choice of coefficient vector norms to be regularized determines whether a Lasso-, Ridge-, or ElasticNet-regularized model is estimated, for the same functional form.<sup>9</sup> We are particularly interested in Lasso-regularized models, since the Lasso penalization procedure encourages zero entries in the coefficient vectors.

<sup>8</sup>Earlier contributions include Menzly, Santos, and Veronesi (2004), where the CCAPM features both time-varying risk preferences and expectations of dividend growth. Santos and Veronesi's (2006) CAPM conditioning variables depend on the shares of the firms' dividends and wages over consumption.

<sup>9</sup>Note that the ML techniques that we use to regularize the coefficients allow us to maintain the same linear functional form, and the convexity of the problem allows us to apply well-known optimization procedures. Refer to the Supplementary Material for further details of the precise regularization schemes employed.



For each ML estimation procedure, the extent of penalization/regularization applied to some coefficient vector  $\theta$  is determined by the value of the corresponding hyperparameter(s)  $\lambda$ .

We evaluate all our models OOS: Accordingly, we adopt a training, validation, and testing procedure intended to avoid overfitting.

### III. Methodology

We present the predictive models in [Section III.A](#), and develop an empirical framework to benchmark our models by their OOS predictive accuracy and test their statistical significance in [Section III.B](#). Finally, [Section III.C](#) describes two grouping procedures: industry membership and  $k$ -means clustering.

#### A. Predictive Models

The approach taken by existing studies of firm-level return predictability is to estimate a predictive model on all firms: For example, GKX apply a variety of linear and nonlinear models, but each model is estimated on all available firms.<sup>10</sup> In this article, we call all such models “pooled models.” A pooled model predicts a firm  $i$ ’s next-period excess returns  $r_{i,t+1}$  based on an  $M \times 1$  vector of current characteristic values  $\mathbf{c}_{it}$  using the same set of estimated coefficients: an intercept  $\alpha_0$  and an  $M \times 1$  vector of pooled coefficients  $\theta_0$ . We write it first in scalar and then in vector form, as follows:

$$(4) \quad r_{i,t+1} = \alpha_0 + \overbrace{\theta_{10}c_{it}^{(1)} + \theta_{20}c_{it}^{(2)} + \dots + \theta_{M0}c_{it}^{(M)}}^{\text{pooled set of coefficients}},$$

$$(5) \quad = \alpha_0 + \theta_0' \mathbf{c}_{it}. \quad (\text{pooled model})$$

Pooled models can range in complexity from an intercept only (i.e., a pooled mean) to more sophisticated regularized models such as the ElasticNet of Zou and Hastie (2005).

We diverge from the existing predictability literature by conjecturing that firms can be partitioned into  $K$  groups. We use a vector of predictive coefficients  $\theta_j$  associated with  $z_i$ , which is the group  $j$  that a firm  $i$  belongs to. We write this first in scalar then in vector notation, using  $M \times 1$  vectors  $\theta_j$  and  $\mathbf{c}_{it}$ :

$$(6) \quad r_{i,t+1} = \alpha_j + \overbrace{\theta_{1j}c_{it}^{(1)} + \theta_{2j}c_{it}^{(2)} + \dots + \theta_{Mj}c_{it}^{(M)}}^{\text{group } j\text{-specific set of coefficients}},$$

$$(7) \quad = \alpha_j + \theta_j' \mathbf{c}_{it}.$$

<sup>10</sup>Likewise, Han et al. (2021) may use different forecasting models (one per characteristic), but each of these models is also trained on a cross section of available firms before the multiple model forecasts (each produced for the entire cross section of firms) are aggregated together in a final step. Their forecast combination approach thus consists in aggregating the predictions of multiple “pooled” models.



In order to predict the excess returns of *any* firm  $i$ , we first estimate  $K$  by-group models, one for each of the  $K$  groupings of firms, and then produce forecasts using the model associated with firm  $i$ 's group  $j$ . In this way, predictive relationships are heterogeneous across the  $K$  firm groupings. Using a scalar indicator variable  $\mathbb{1}\{z_i = j\}$  to denote firm  $i$  being a member of group  $j$ , we combine these  $K$  models as follows:

$$(8) \quad r_{i,t+1} = \sum_{j=1}^K \left( \alpha_j + \boldsymbol{\theta}'_j \mathbf{c}_{it} \right) \mathbb{1}\{z_i = j\}. \quad (\text{by-group model})$$

Note that each coefficient vector  $\boldsymbol{\theta}_j$  must be regularized during the estimation procedure, and the degree of penalization is controlled by a group-specific hyperparameter  $\lambda_j$ , of which there are therefore  $K$  in total for the Lasso- and Ridge-regularized variants. Each group-specific parameter is tuned separately, for the subset of firms belonging to the group.

We also specify and evaluate a third class of (hybrid) models that melds the pooled specification with the by-group specification. We continue to hypothesize that the conditional mean function varies across (groups of) firms in the cross section while now explicitly allowing a set of coefficients to be shared across all firm groupings: This permits each estimated coefficient to be explicitly interpreted as pertaining to all firms in the cross section or solely a specific group within the cross section of firms. Both sets of coefficients are estimated in two stages: In the first stage, we estimate coefficients that are common to all groups (the pooled set) and do so on all available ("pooled") samples:

$$(9) \quad r_{i,t+1} = \overbrace{\alpha_0 + \theta_{10}c_{it}^{(1)} + \theta_{20}c_{it}^{(2)} + \dots + \theta_{M0}c_{it}^{(M)}}^{\text{pooled set of coefficients}},$$

$$(10) \quad = \alpha_0 + \boldsymbol{\theta}'_0 \mathbf{c}_{it}. \quad (\text{stage 1, pooled})$$

We then produce predictions  $\hat{r}_{i,t+1}$  for every element in the pooled sample. Using the known values  $r_{i,t+1}$ , we calculate prediction residuals. These residuals are now the inputs to the second stage. The second stage estimates one model for each of the  $K$  (nonoverlapping) groups of firms, each indexed  $j$ :

$$(11) \quad \underbrace{(r_{i,t+1} - \hat{r}_{i,t+1})}_{\text{first-stage residuals}} = \overbrace{\alpha_j + \theta_{1j}c_{it}^{(1)} + \theta_{2j}c_{it}^{(2)} + \dots + \theta_{Mj}c_{it}^{(M)}}^{\text{group } j\text{-specific set of coefficients}},$$

$$(12) \quad = \alpha_j + \boldsymbol{\theta}'_j \mathbf{c}_{it}. \quad (\text{stage 2, by-group})$$

Predictions for firms' excess returns are thus the sums of the forecasts from each stage.<sup>11</sup> The first stage of the model should estimate predictive relationships that are common to all firms in the cross section, whereas the second stage should

<sup>11</sup>Each firm is associated with two models: The first-stage model is shared with all other firms, and the second-stage model is shared with other firms in the same group.

detect any residual predictive relationships that are specific to groups of firms. The two-stage model is linear,

$$(13) \quad r_{i,t+1} = \underbrace{\alpha_0 + \theta'_0 \mathbf{c}_{it}}_{\text{pooled stage}} + \underbrace{\sum_{j=1}^K \left( \alpha_j + \theta'_j \mathbf{c}_{it} \right) \mathbb{1}\{z_i = j\}}_{\text{by-group stage}}, \quad (\text{two-stage model})$$

and we evaluate variants that have been regularized using Lasso and Ridge penalties. As in the by-group model, this requires multiple hyperparameters:<sup>12</sup> in this case,  $K + 1$  hyperparameters  $\lambda_0, \lambda_1, \dots, \lambda_K$  to control the degree of regularization of the  $K + 1$  coefficient vectors  $\theta_0, \theta_1, \dots, \theta_K$ . Note that if pooled models fail to detect some group-specific predictive relationships that are successfully captured by the second stage of our two-stage procedure, then we would expect the overall two-stage model to outperform pooled models OOS.

Finally, note that each of the individual pooled or group-specific models in our study is a linear function of characteristic inputs  $\mathbf{c}_{it}$ . Therefore, since each firm  $i$  belongs to a single group  $j$ , the same is true for the composite by-group and two-stage models when the group-membership variables  $z_i$  are fixed/prespecified, such as when they are defined by industry membership. Nonlinearities can arise only when the  $z_i$  are functions of the characteristics: This is the case when  $k$ -means clustering is used to infer the partition of firms.

## B. Estimation, Performance, and Tests

To produce OOS performance estimates, we must hold out a test set of samples upon which to evaluate our models; in addition, the need to tune regularization hyperparameters requires us to also prepare a validation set of samples for that purpose. Therefore, for a fixed time period, we partition our data into three disjoint subsets: for training, validation, and testing. Following GKX, we split our database into multiple temporal *slices*, with the training set growing by 1 year with each slice and the subsequent validation and test sets shifting forward by 1 year each and maintaining a constant size. There are 6 slices in total. Since we have a shorter overall sample than GKX do, we have shortened each of the 3 sets in comparison: The validation set is always 6 years in length, and the test set in each slice is 1-year long; so the test set in slice 1 is 2010, and the test set in slice 6 is 2015. The sequencing of the training set, validation set, and test set take the time-ordered nature of the returns data into account: It is important for the training set to precede the validation set, and the validation set to precede the test set, in order to preserve the temporal ordering of the data.<sup>13</sup>

Our tuning procedure is as follows: In each slice, any given model that requires tuning of some hyperparameter(s)  $\lambda$  is estimated multiple times on the training set

<sup>12</sup>We omit ElasticNet-regularized variants of the two-stage model as these would require twice as many hyperparameters to be tuned as Lasso- and Ridge-regularized variants.

<sup>13</sup>This explains why we cannot reorder the training, validation, and test sets with classical cross-validation.

for a range of possible  $\lambda$  values. Then the optimal value of each  $\lambda$  is selected based on the performance of the model on the validation set, as measured by mean-squared error (MSE). Finally, the entire model is re-estimated on a concatenation of the training and validation sets, for the optimal value of each  $\lambda$ . Any models that do not require tuning of some  $\lambda$  hyperparameter(s) are directly estimated on the concatenation of the training and validation sets. This procedure enables us to assess each model's OOS performance on the test set, per slice. We report OOS performance figures per model, computed on all its test set forecasts.

This empirical design produces true OOS estimates of predictive performance, since no model is estimated in any way on any data within the test set, only assessed on these data. GKX and Gu et al. (2021) report model performance using the following OOS predictive  $R^2$ , and we follow those studies in doing likewise:

$$R_{\text{OOS}}^2 = 1 - \frac{\sum_{i,t} (r_{i,t+1} - \hat{r}_{i,t+1})^2}{\sum_{i,t} r_{i,t+1}^2}.$$

This metric is equivalent to the classical estimator for the fraction of variance explained  $R^2$  without demeaning the denominator; alternatively, we can understand it as incorporating the assumption of a zero mean in the denominator, or as benchmarking our models against forecast values of zero.

Comparing  $R_{\text{OOS}}^2$  values is complicated by the absence of suitable parametric tests involving this metric. While the  $R^2$  is closely linked to the MSE, testing for differences in MSEs would require additional assumptions for our high-dimensional cross-sectional setting: For example, linking to a likelihood ratio test in the style of Lien and Vuong (1986) is only suitable for a low-dimensional Gaussian linear model. Diebold and Mariano (1995)-style tests also involve MSE comparisons; however, Timmermann (2018) argues that such tests have limited power to detect return predictability because of the prevalence of weak predictors in our asset-pricing context. Furthermore, members of the Diebold and Mariano (1995) family of tests compare time series of predictions, and we would like to compare cross sections of predictions.

Accordingly, we rely on a nonparametric bootstrap analysis to perform statistical tests directly on our  $R_{\text{OOS}}^2$  quantities of interest. Since our study focuses on the cross section of firms, each bootstrap sample consists of a set of firms (permnos) drawn from the cross section of the full sample of firms, with replacement. This resampling scheme is also used by “panel bootstraps” in a regression setting (see Cameron and Trivedi ((2005), p. 377)) or Kapetanios (2008)). To avoid any potential bias, we take care to draw the full time series of returns and characteristics for each firm (permno) across data slices while preserving the temporal structure of each panel sample. Given each sampled panel, we fully estimate (including hyperparameter turning) any models that we wish to compare, then generate predictions, from which we calculate the appropriate  $R_{\text{OOS}}^2$  quantities. In comparing two models, we are interested in the difference between the corresponding pair of  $R_{\text{OOS}}^2$  values, and this difference in values is the bootstrap statistic of interest for such a comparison. Using multiple bootstrap samples of this quantity,

we may then calculate bootstrap confidence intervals as part of a statistical test of the  $R^2_{\text{OOS}}$  difference of interest.<sup>14</sup>

Note that, although this procedure is computationally expensive, it is achievable in practice thanks to the convex objective functions and efficient estimation procedures of the ML models that we use in this study; the use of costlier methods (such as neural networks) would have hindered such an analysis.

Re-estimating our models on each bootstrap sample also allows us to compute bootstrap confidence intervals for each estimated coefficient. We use these to test individual hypotheses on coefficient values differing from 0 for any given significance level, much like the analysis that typically follows a classical linear regression.<sup>15</sup>

### C. Partitioning the Cross Section of Firms

To illustrate our results, we consider two methods for partitioning the cross section of firms into groups of related firms. There are, of course, a large number of arbitrary firm partitions that are possible. First, and given our earlier literature review, we use industry classifications.<sup>16</sup> Then we describe a partitioning criterion that effectively makes uncovering heterogeneity more challenging as it infers group memberships directly from the data using a multidimensional metric.

#### 1. SIC Codes

We define firm groupings based on industry classifications when estimating by-group predictive models.

A firm's industry is based on its SIC code, according to the ranges defined in [Table 1](#). SIC codes are qualitative labels assigned by the U.S. government according to the nature of firms' primary business activities at the time of assignment. If they are accurate,<sup>17</sup> then firms belonging to the same SIC code range should engage in similar activities. We do not assign more granular industry classifications beyond the standard high-level groupings defined in [Table 1](#) because that would require us

<sup>14</sup>The nonparametric bootstrap makes minimal distributional assumptions. Furthermore, our application is a particularly suitable one because our quantity of interest, the incremental  $R^2_{\text{OOS}}$ , does not fall under the special cases that can lead to inconsistencies in the bootstrap inference (Chernick (2007), Ch. 9); For example, it is not an extremal statistic, nor is it generated from a distribution with nonexistent moments. Above all, as mentioned earlier, we take into account potential time series dependencies in the sampling process.

<sup>15</sup>Interestingly, exact post-selection statistical tests have recently been developed for Lasso-regularized models, like some of our own, notably the "fixed- $\lambda$ " test of Lee, Sun, Sun, and Taylor (2016) and the "spacing test" of Tibshirani, Taylor, Lockhart, and Tibshirani (2016). However, we found it impractical to perform the computations necessary for those two exact tests on our large data set, for the purposes of comparing different estimates. Indeed, Lee et al. ((2016), p. 921) discuss one such practical limitation to their test. In contrast, the bootstrap, though computationally expensive, can be run without these practical issues, and is also general enough to produce confidence intervals for ML models that have been regularized with penalties other than the Lasso.

<sup>16</sup>Our baseline industry partition is based on SIC codes. In Appendix C of the Supplementary Material, we consider an alternative industry partition based on Hoberg and Phillips's (2016) Text-Based Network Industry Classifications.

<sup>17</sup>Since SIC codes are manually assigned and rarely changed, it is possible that they do not accurately reflect firms' activities.

TABLE 1  
Industries

Table 1 reports industry groupings, as determined by firms' SIC codes.

Industry	SIC Code Range
Agriculture	0100–0900
Construction	1520–1731
Finance	6020–6799
Manufacturing	2000–3990
Mining	1000–1400
Noclassif	9995–9997
Retail	5200–5990
Services	7000–8900
Transport and Utilities	4011–4991
Wholesale	5000–5190

to make a further subjective decision on what level of grouping in the hierarchical SIC code structure would be most useful.

2. Cluster Analysis

Given our focus on firms and their characteristics, we turn to an ML technique that allows us to use these observable characteristics to infer groupings of firms. *Cluster analysis* is a statistical and ML technique that aims to partition data points into clusters, and the *k*-means algorithm is one of the most popular ways of doing so (see Hastie, Tibshirani, and Friedman (2009)). We use *k*-means clustering to group firms into clusters of similar firms, where similarity is measured by (squared) Euclidean distances between (the means of) these observable characteristics.

As a clustering technique, *k*-means is nonparametric, freeing us from making distributional assumptions in this high-dimensional setting. It requires only one hyperparameter to be specified: the number of clusters (which we use a well-established criterion to do). Furthermore, each characteristic is considered and is weighted equally during the *k*-means cluster formation procedure. We therefore define a transparent benchmark that summarizes the full high-dimensional set of characteristics.<sup>18</sup>

In our setting, each input point  $\mathbf{x}_i$  to the clustering algorithm consists of the scaled means of firm  $i$ 's characteristics vector  $\mathbf{c}_{it}$ ; the elements of  $\mathbf{x}_i$  do not include firm  $i$ 's industry nor do they include the firm's excess return. For each firm  $i$ , the procedure outputs a scalar  $\gamma_i$  that denotes the firm's assignment to a cluster (i.e., its estimated latent group membership).

We provide further details of the *k*-means algorithm in the Supplementary Material. The number of clusters to use in each data set  $K$  is a free parameter to the *k*-means clustering algorithm. We use the silhouette technique of Rousseeuw (1987) to pick the optimal  $K^*$  number of clusters. Since we are using the clusters as inputs to predictive models, the OOS performance of those models may suffer if we pick an incorrect number of clusters; we will see in the results that this

<sup>18</sup>As well as the benefit of transparency, the levels of predictability that we detect in our study are likely to be a conservative benchmark if more sophisticated grouping techniques can be applied by future research.

method of picking  $K^*$  does not appear to be too simple that it interferes with good predictive performance.

The clustering procedure is performed at the firm level. This requires a choice on how to deal with firms that drop into and out of the samples: For each slice, we perform the  $k$ -means clustering procedure on the set of firms that are present in both the training and validation sets, avoiding the possibility that a small cluster is formed based on firms that later drop out of the sample. Similarly, we do not produce predictions for any firms that first appear in each slice's (year-long) test set, as the firm is not assigned to a cluster and we do not wish to make an arbitrary cluster assignment choice.

Note that  $k$ -means clustering is not provided with any mapping from a firm  $i$  to its assigned cluster  $\gamma_i$ ; rather, it aims to infer the  $\gamma_i$  cluster memberships based on observables  $\mathbf{x}_i$ . We therefore assume that the observables in our data set can be used to proxy for firms' latent group memberships. In passing cluster assignments  $\gamma_i$  to our predictive models, we are effectively combining an unsupervised learning stage with a supervised learning one.<sup>19</sup>

## IV. Data

### A. Databases

In building our database of firm-level data, we begin with the CRSP returns of firms quoted on the major U.S. exchanges. We follow the literature in our construction of the CRSP database: We consider share classes 10 and 11, and NYSE, AMEX, and Nasdaq-listed firms. Microcap stocks (i.e., with a market capitalization of \$100 million or less) and illiquid stocks (i.e., with a monthly traded volume of less than \$100K) are excluded. Furthermore, banking and utility stocks (from Kenneth French's website) are excluded.

We join this CRSP database to Compustat and IBES in order to prepare firm-level data that we use to construct firm-level characteristics.

Our database begins in 1980, as this is when most firm characteristics become widely available, and ends in 2015.

### B. Characteristics

Full details of all variables used in this study can be found in Tables IA.1 and IA.2 in the Supplementary Material.

We follow GHZ in defining firm-level characteristics from the CRSP, Compustat, and IBES databases. The firm characteristics in question are also known in the literature as "anomaly characteristics" because portfolios formed by sorting on such

<sup>19</sup>In this sense, our application of  $k$ -means clustering to observable variables can be distinguished from the models of Bonhomme and Manresa (2015), which include variants with cluster-varying fixed effects and cluster-varying coefficients. The models introduced by that study require joint estimation of the regression and clustering steps, so cluster assignments are made based on their contributions toward minimizing the sum of squared residuals of the predicted outcome variables (i.e., firm-level excess returns, in our case) rather than the predictive variables themselves (i.e., characteristics, in our case). GKX found that OLS-based methods perform poorly in a comparable high-dimensional setting to ours, so the models of Bonhomme and Manresa (2015) would not be appropriate for that reason.

characteristics have been found to result in anomalous excess returns. We have collected 101 such firm-level characteristics, in common with the 102 that were collected by GHZ.<sup>20</sup>

We have also incorporated a further 8 market-level variables, as provided by Welch and Goyal (2007): BM\_MKT, DFY\_MKT, DP\_MKT, EP\_MKT, NTIS\_MKT, SVAR\_MKT, TBL\_MKT, and TMS\_MKT. This takes us to a total of 109 predictive variables. Note that we have appended “\_MKT” to the names of all market-level variables in order to distinguish them from firm-level characteristics.

In relation to GKK, we use more firm-level characteristics. We omit REALESTATE, which GKK included. We include CHFEPS, CHNANALYST, DISP, FGR5YR, IPO, NANALYST, SFE, and SUE, which GKK omitted, and we use the same set of market-level variables, without any interaction terms.<sup>21</sup>

### C. Preprocessing

We follow GKK in a number of key data preprocessing steps. Firm-level characteristics are rescaled to the range  $[-1, +1]$  cross-sectionally, that is, per month. This does not apply to market-level variables. Any missing values for firm-level characteristics are replaced with the cross-sectional medians. No winsorization or other form of trimming is applied to the data.

Following the literature, firm-level characteristics that vary annually are lagged by 6 months, and those that vary quarterly are lagged by 4 months. This is done to mitigate any potential look-ahead bias, since these characteristics are typically made public with a delay.

Although our database begins in 1980, our study uses characteristics from 1984 onward. This is for two reasons: The lagging procedure described above shifts forward the start date, and some firm-level characteristics require prior data in their construction,<sup>22</sup> shifting the start date forward even further. Starting from 1984 means all characteristics are available.

## V. Empirical Results

### A. Clusters

As discussed in our methodology section, we apply the  $k$ -means algorithm to observable characteristics. The algorithm is applied to the training set within each slice, so that the hyperparameter tuning step, which uses the validation set, selects among a set of fully specified models. Our method of selecting the optimal number of clusters consistently returns 2–4 clusters, no matter whether we apply

<sup>20</sup>We omit the REALESTATE firm-level characteristic, which GHZ included, because our sample did not include any observations prior to 1985.

<sup>21</sup>Refer to the Supplementary Material for details of how the full set of firm-level and market-level variables have been constructed. No other predictive variables are used in this study. One of our contributions is to push forward the usefulness and applicability of linear-based models by showing that they are capable of detecting substantial levels of OOS firm-level return predictability; including nonlinearities and interactions would obscure this contribution.

<sup>22</sup>GRAPX requires 2 years of prior data. BETA and IDIOVOL each require 3 years of prior returns.



it to the top 1,000 (by market capitalization), the top 2,000, or all available firms in our sample.

The resulting clusters are depicted in [Figure 1](#): Each point represents a single firm, with its high-dimensional characteristics collapsed into 2 dimensions based on the first 2 principal components of the characteristics data. Interestingly, the firm groupings are apparent even in such high dimensions, and even though we do not apply any dimensionality reduction procedure before the  $k$ -means clustering step. Visually, the clusters appear reasonable in principal component space: Firms appear close together to the naked eye, and the relative spatial distribution of firms does not change much from slice to slice. This stability of the inferred clusters is a welcome feature when interpreting these clusters in terms of characteristics.<sup>23</sup> It also facilitates comparisons between industry partitions and cluster partitions, since we do not unfairly compare almost-static industry memberships to fast-moving cluster memberships, as both are stable in our study.

### 1. Interpreting the Clusters

We examine the characteristics of the firms that make up the various clusters during the cluster formation process.

[Figure 2](#) plots differences of (cross-sectional) characteristic means by cluster versus the means across the remaining clusters. Intuitively, the larger any one difference (represented by a bar) is, the more this characteristic stands out for the given cluster when compared to all the other clusters. All firm-level characteristics are ordered by the largest such difference to the smallest, calculated for the sixth and final slice in order to maintain a consistent ordering throughout the figure. The top 20 such firm-level characteristics are displayed.

Consistent with [Figure 1](#), the cluster compositions in [Figure 2](#) are also stable across slices. The exception is the detection of the fourth cluster in the sixth and final slice, which appears to be concentrated among sin stocks. Based on the cross-sectional mean characteristic differences among clusters in [Figure 2](#), we characterize the 3 stable clusters as follows:

- Cluster 1 is comprised of older firms, with relatively low analyst earnings forecasts, a lower likelihood of secured debt, and relatively high operating profitability and sales growth compared to inventory growth.
- Cluster 2 is comprised of younger firms that are relatively cash-poor and whose earnings surprises are relatively lower (and more negative).
- Cluster 3 is concentrated among very young firms (such as the recently IPO'ed) and those with relatively poor profitability.

These results indicate that making use of only a few clusters has enabled us to partition firms according to interpretable economic criteria. It is worth clarifying that since clusters are defined based on the means of firms' characteristics during the training period, these characteristics will continue to evolve during the subsequent validation and test periods. Cluster memberships during later periods should

<sup>23</sup>Table IA.4 in the Supplementary Material measures cluster stability based on the fractions of firms that remain in each cluster during slice transitions. Firm cluster memberships are highly stable according to this criterion.

FIGURE 1  
Cluster Visualizations

Figure 1 shows the visualizations of the learned clusters, per slice. Each point represents a single firm. The x- and y-axes represent the first and second principal components, respectively, of the firms' mean characteristics. All firms in our sample are represented.

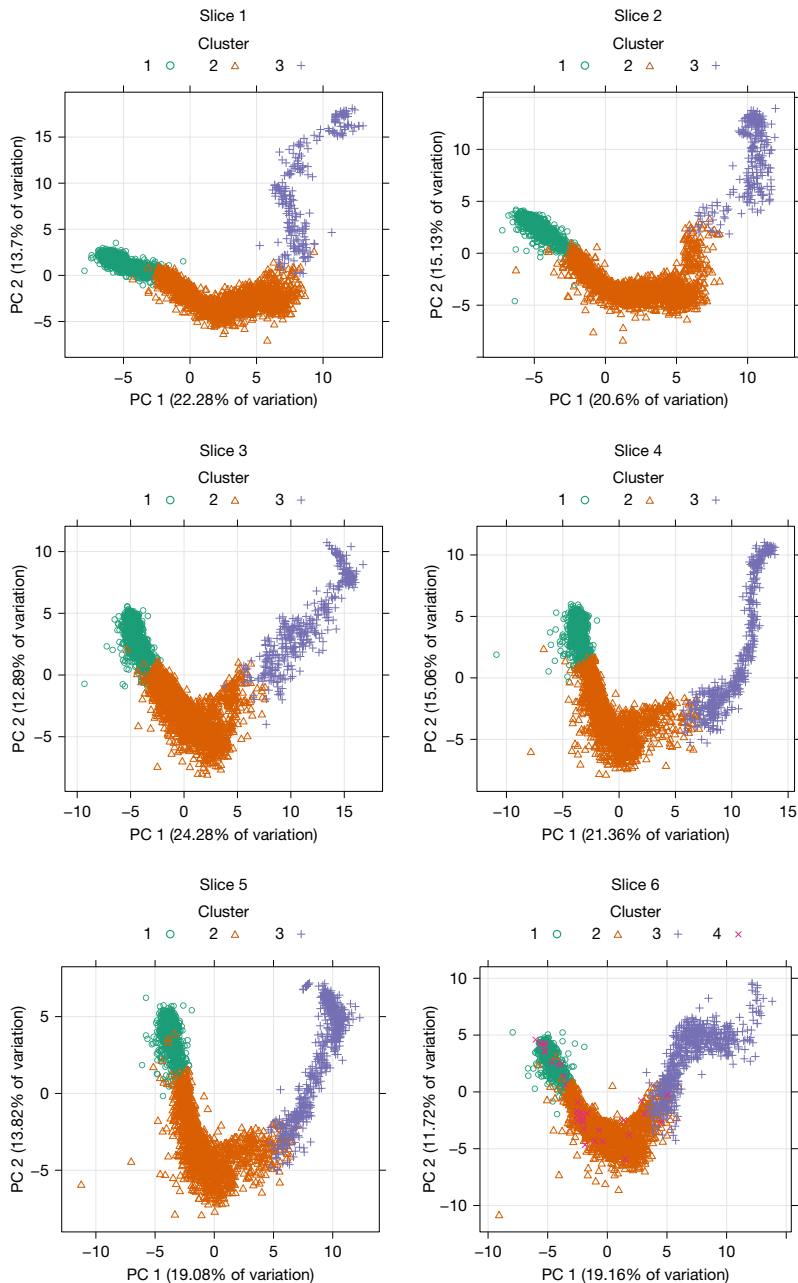
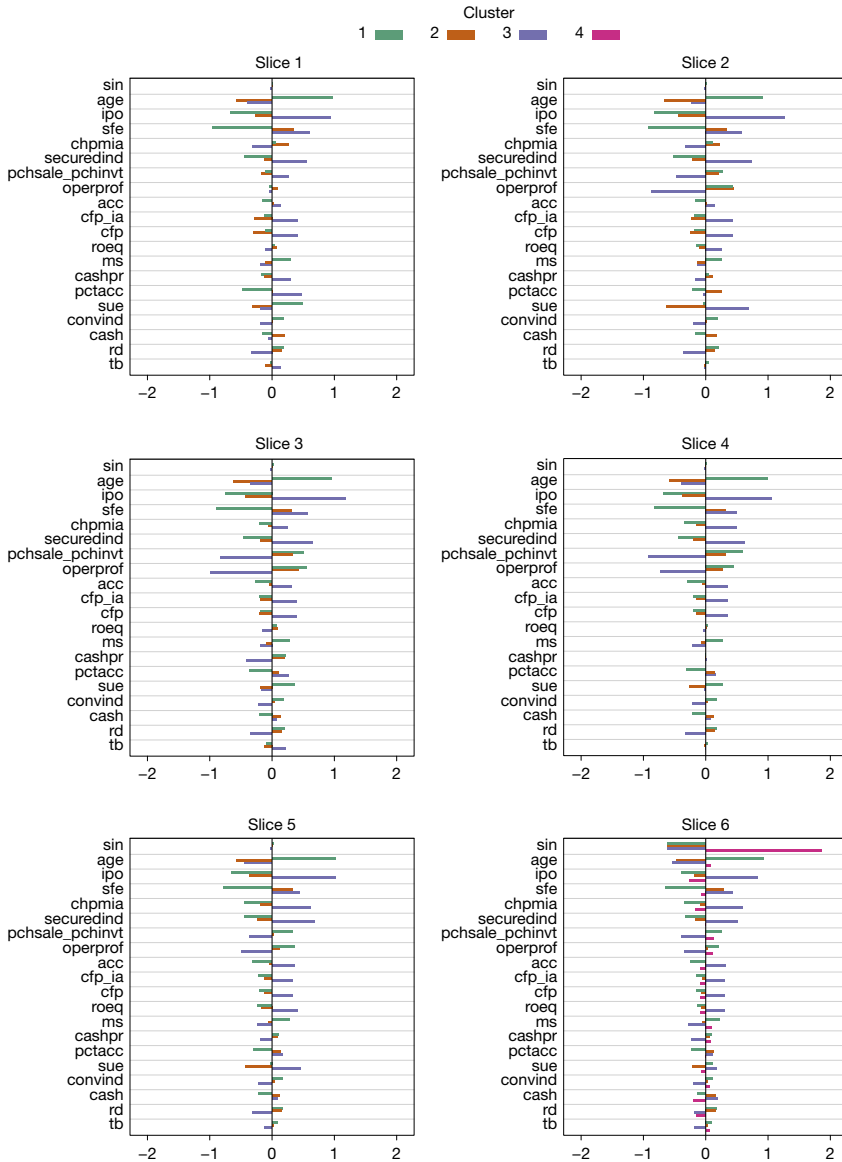


FIGURE 2  
Cluster Interpretations

In Figure 2, we interpret clusters by comparing their characteristic means. Each bar depicts the difference between a cluster's (cross-sectional) characteristic mean and the characteristic mean calculated for all other clusters. The 20 firm-level characteristics with the highest such absolute deviations (for any one cluster) are shown, in order of the highest deviation in the sixth and final slice (top) to the lowest deviation (bottom). Each pane visualizes a single slice's clusters.



Difference in cross-sectional characteristic means between one focal cluster and the remaining non-focal clusters

be interpreted accordingly; for example, no firms will be recently IPO'd during the OOS test period.

## 2. Comparison with Existing Results

We are not aware of prior attempts to cluster firm-level characteristics by their cross-sectional values for the purpose of cross-sectional return prediction. In a recent study, Balasubramaniam et al. (2023) clustered (Indian) firm-level characteristics by investor holdings and found that firm age is an important characteristic for explaining investors' holding patterns. It is notable that firm age is one of the most important characteristics of cluster formation in our setting. We will show later that the characteristics that matter for return prediction vary across clusters that were formed (in part) based on a characteristic that Balasubramaniam et al. (2023) found "has the strongest investor clientele," but that firm age itself is not selected as a direct predictor of returns. Our evidence, therefore, suggests a role for how investor clienteles may impact return predictability.<sup>24</sup>

## B. Predictive Performance and Heterogeneity

Our first result on firm-level return predictability is that exploiting an economic partition of the cross section of firms can positively impact OOS predictability. To show this, we use the industry partition of the cross section to estimate pooled models, as defined in equation (5), as well as by-industry models, as defined in equation (8), in order to compare their OOS predictive performance. We employ a variety of regularization schemes to deal with the challenging nature of this high-dimensional data set, and also report results from unregularized OLS models.

Panel A of Table 2 reports the attained OOS predictability results by industry. Other than unregularized OLS models, the OOS predictive performance appears to be positive across the board, suggesting that by-industry and pooled specifications can successfully detect conditional risk premia. We now turn to a comparison between specifications.

Panel B of Table 2 reports the *incremental* OOS predictability of by-industry over pooled models; each cell contains the  $R^2_{\text{OOS}}$  (%) of a by-group model minus the  $R^2_{\text{OOS}}$  (%) of the pooled model with the same regularization scheme (Lasso, ElasticNet, or Ridge), computed per industry. These values should be positive if heterogeneity positively impacts predictability. Other than the extractive industries (agriculture and mining), the incremental OOS predictability values do appear positive across the board. Our statistical test is based on confidence intervals computed from 1,000 bootstrap samples; recall that each bootstrap sample includes re-running the estimation, tuning, and prediction steps for all these models. Importantly, all the values that our bootstrap procedure concludes are significantly

<sup>24</sup>Idiosyncratic volatility is an example of a characteristic that Dorn and Huberman (2010) highlighted in prior work on investor clienteles. While this characteristic does not stand out in isolation as important for determining cluster membership, we note that state variables related to age, size, and profit margin that have a bearing on explaining trends in idiosyncratic volatility (Brown and Kapadia (2007)) are important for cluster formation. IDIOVOL variation may be (indirectly) captured by these membership-determining characteristics.

TABLE 2  
Industry-Level Predictability

Table 2 reports industry-level out-of-sample predictability, measured by  $R^2_{OOS}$  (%), when partitioning firms by industry. Panel A reports the  $R^2_{OOS}$  predictive performance results. Panel B reports the *incremental*  $R^2_{OOS}$  by regularization scheme, calculated as the difference between by-industry and pooled  $R^2_{OOS}$ , together with hypothesis test results: Asterisks in Panel B indicate that the increments are significantly different from 0, based on bootstrap confidence intervals. The models are each estimated based on all available firms, once per slice. The significance tests in Panel B are based on 1,000 bootstrap samples. Only results for the regularized linear models are reported in Panel B, since OLS-based by-industry models were not invertible for multiple bootstrap realizations of the high-dimensional data set. \*\*\*, \*\*, and \* denote significance at the 99%, 95%, and 90% levels, respectively.

Panel A.  $R^2_{OOS}$

Model	Agriculture	Construction	Finance	Manufacturing	Mining	Noclassif	Retail	Services	Transport and Utilities	Wholesale
By-industry	0.82	1.40	0.65	0.73	0.34	1.59	0.53	0.70	0.61	0.86
Lasso										
By-industry	0.70	1.60	0.76	0.75	0.31	1.94	0.46	0.80	0.69	0.97
ElasticNet										
By-industry	0.83	1.61	0.81	0.78	0.32	1.97	0.48	0.85	0.74	1.02
Ridge										
By-industry	-9.26	-6.58	-5.99	-4.79	-6.97	0.91	-5.58	-5.62	-3.24	-5.24
OLS										
Pooled Lasso	1.02	0.98	0.37	0.36	0.47	0.34	0.25	0.32	0.28	0.36
Pooled	1.05	1.04	0.32	0.26	0.44	0.22	0.10	0.25	0.17	0.31
ElasticNet										
Pooled Ridge	1.05	1.04	0.32	0.26	0.44	0.22	0.10	0.25	0.17	0.31
Pooled OLS	-3.55	-2.81	-4.14	-5.30	-5.82	-3.73	-6.53	-4.08	-4.81	-4.69

Panel B. Incremental  $R^2_{OOS}$  with Hypothesis Test Results

Regularization	Agriculture	Construction	Finance	Manufacturing	Mining	Noclassif	Retail	Services	Transport and Utilities	Wholesale
Lasso	-0.20	0.42	0.28	0.37***	-0.13	1.25	0.28	0.38**	0.33	0.50
ElasticNet	-0.35	0.56	0.44	0.49***	-0.13	1.72	0.36	0.55***	0.52	0.66
Ridge	-0.22	0.57	0.49	0.52***	-0.12	1.75	0.38	0.60	0.57	0.71

different from 0 and are also positive. This statistical evidence indicates that allowing heterogeneous predictive relationships across industries can positively impact OOS predictability at the firm level.

We saw in Section V.A that applying the  $k$ -means clustering procedure to our cross section of firms resulted in economically interpretable groupings. We now consider whether this partition of clusters can also be used to improve OOS predictability at the firm level.

Panel A of Table 3 reports the attained OOS predictability results by cluster. Once again, and besides the unregularized OLS models, the OOS predictive performance appears to be positive across the board, suggesting that by-cluster specifications can also successfully detect conditional risk premia in the cross section of firms.

Panel B of Table 3 reports the *incremental* OOS predictability of by-cluster over pooled models; each cell contains the  $R^2_{OOS}$  (%) of a by-cluster model minus the  $R^2_{OOS}$  (%) of the pooled model with the same regularization scheme, computed per cluster. Once more, for heterogeneity to positively impact predictability, these values should be positive, and this appears to be the case for stable clusters and the Lasso and ElasticNet regularization schemes. To perform statistical tests, we repeat our earlier bootstrapping procedure to produce confidence intervals computed from 1,000 bootstrap samples: We find that most values in the table are significantly different from 0. Therefore, for Lasso- and ElasticNet-regularized models, this statistical evidence indicates that allowing heterogeneous predictive relationships across clusters can also positively impact OOS predictability at the firm level.

TABLE 3  
Cluster-Level Predictability

Table 3 reports cluster-level out-of-sample predictability, measured by  $R^2_{OOS}$  (%), when partitioning firms by cluster. Panel A reports the  $R^2_{OOS}$  predictive performance results. Panel B reports the incremental  $R^2_{OOS}$  by regularization scheme, calculated as the difference between by-cluster and pooled  $R^2_{OOS}$ , together with hypothesis test results: Asterisks in Panel B indicate that the increments are significantly different from 0, based on bootstrap confidence intervals. Cluster 4 is omitted because it is only estimated for a single slice, and therefore an  $R^2_{OOS}$  figure cannot be produced. The models are each estimated based on all available firms, once per slice. The significance tests in Panel B are based on 1,000 bootstrap samples. Only results for the regularized linear models are reported in Panel B, since OLS-based by-cluster models were not invertible for multiple bootstrap realizations of the high-dimensional data set. \*\*\*, \*\*, and \* denote significance at the 99%, 95%, and 90% levels, respectively.

Panel A. $R^2_{OOS}$			
Model	Cluster 1	Cluster 2	Cluster 3
By-cluster Lasso	1.09	1.08	0.74
By-cluster ElasticNet	1.08	1.09	0.70
By-cluster Ridge	0.99	1.03	0.59
By-cluster OLS	-64.11	-60.96	-54.68
Pooled Lasso	1.05	1.01	0.59
Pooled ElasticNet	1.02	0.97	0.64
Pooled Ridge	1.03	0.98	0.65
Pooled OLS	-4.74	-4.94	-4.50
Panel B. Incremental $R^2_{OOS}$ with Hypothesis Test Results			
Regularization	Cluster 1	Cluster 2	Cluster 3
Lasso	0.04*	0.07*	0.15***
ElasticNet	0.06*	0.12*	0.06***
Ridge	-0.04***	0.05	-0.06***

C. Predictive Performance with Two-Stage Procedure

Following the prior literature on firm-level return predictability, we now assess OOS predictive performance on the cross section of firms without conditioning by industry or by cluster. In doing so, we follow the procedure laid out in equations (9)–(13) to incorporate two-stage models.

We train our predictive models on 3 subsets of the data, based on rankings of firm market capitalizations. The first subset (“top 1,000”) consists of the largest 1,000 firms by market capitalization, the second subset (“top 2,000”) consists of the largest 2,000 firms by market capitalization, and the final subset does not discard any firms based on their size. We then evaluate each trained model’s predictions on the corresponding set of firms by market cap; we condition on size in this manner to ensure that our findings are not driven by the smallest firms.

Table 4 reports the aggregate OOS predictive performance of each model that we consider, when taking SIC codes as the grouping criterion. The models range from simpler pooled models to more sophisticated ones that exploit the industry partition of the cross section. Each panel reports results from models that have been trained and tested on the various subsets of firms: It is thus more meaningful to compare results within the table panels rather than across them.

Many of the models exhibit a positive OOS  $R^2_{OOS}$ , in spite of the difficulty of predicting firm-level returns. The best-performing model in each case is a two-stage Ridge-regularized model. Ridge regression models select all available characteristics and are most suited to situations where a large number of correlated predictive variables are present, as they shrink such coefficients toward one another (Hastie et al. (2009)). It is worth noting that not only does OLS perform poorly, but

TABLE 4  
Overall Predictability Using Industries

Table 4 reports aggregate out-of-sample predictability, measured by  $R^2_{OOS}$  (%), when partitioning firms by industry. Each panel represents results from estimating the models based on a particular subset of firms and then generating predictions for that same subset: i) on the largest 1,000 firms by market capitalization, ii) on the largest 2,000 firms, and iii) on the full sample.

Panel A		Panel B		Panel C	
Model	Top 1,000	Model	Top 2,000	Model	All Firms
Two-stage Ridge	1.65	Two-stage Ridge	1.38	Two-stage Ridge	0.76
By-industry Ridge	1.60	By-industry Ridge	1.34	Pooled ElasticNet	0.73
Pooled ElasticNet	1.57	Pooled ElasticNet	1.33	Pooled Ridge	0.73
Pooled Ridge	1.57	Pooled Ridge	1.33	By-industry Ridge	0.72
By-industry ElasticNet	1.54	By-industry ElasticNet	1.31	Pooled Lasso	0.71
Pooled Lasso	1.52	By-industry Lasso	1.29	By-industry ElasticNet	0.69
By-industry Lasso	1.49	Pooled Lasso	1.29	By-industry Lasso	0.65
Two-stage Lasso	1.49	Two-stage Lasso	1.29	Two-stage Lasso	0.65
Pooled OLS	-8.70	Pooled OLS	-7.36	Pooled OLS	-3.77
By-industry OLS	-14.78	By-industry OLS	-12.05	By-industry OLS	-5.44
Two-stage OLS	-14.78	Two-stage OLS	-12.05	Two-stage OLS	-5.44

TABLE 5  
Overall Predictability Using Clusters

Table 5 reports aggregate out-of-sample predictability, measured by  $R^2_{OOS}$  (%), when partitioning firms by cluster. Each panel represents results from estimating the models based on a particular subset of firms and then generating predictions for that same subset: i) on the largest 1,000 firms by market capitalization, ii) on the largest 2,000 firms, and iii) on the full sample.

Panel A		Panel B		Panel C	
Model	Top 1,000	Model	Top 2,000	Model	All Firms
Two-stage Ridge	1.91	By-cluster Lasso	1.61	Two-stage Lasso	1.05
Pooled Ridge	1.91	Two-stage Lasso	1.61	By-cluster Lasso	1.03
Pooled Lasso	1.88	Pooled Lasso	1.60	By-cluster ElasticNet	1.03
By-cluster Ridge	1.86	By-cluster ElasticNet	1.59	Pooled Lasso	0.97
Pooled ElasticNet	1.85	Pooled Ridge	1.58	Two-stage Ridge	0.96
By-cluster ElasticNet	1.83	Two-stage Ridge	1.58	By-cluster Ridge	0.95
Two-stage Lasso	1.78	By-cluster Ridge	1.55	Pooled Ridge	0.95
By-cluster Lasso	1.77	Pooled ElasticNet	1.53	Pooled ElasticNet	0.94
Pooled OLS	-8.86	Pooled OLS	-8.23	Pooled OLS	-4.81
By-cluster OLS	-30.86	By-cluster OLS	-20.92	By-cluster OLS	-61.38
Two-stage OLS	-30.86	Two-stage OLS	-20.92	Two-stage OLS	-61.38

attempting to incorporate heterogeneity in predictive relationships even worsens the predictive performance. This highlights the need for regularization in our high-dimensional setting.

Table 5 reports the aggregate OOS predictive performance when using clusters to partition the cross section of firms. In interpreting these models, we once again focus on within-table comparisons of the OOS  $R^2_{OOS}$  values.

The headline result from Panel C of Table 5 is that the two-stage Lasso-regularized model performs best overall when estimating on and predicting for all available firms. It also exhibits the best individual performance for each of the 3 main clusters. Once again, OLS performs poorly, even when trying to incorporate heterogeneity in predictive relationships.

Compared to the previous industry-grouped predictive results, a key benefit to exploiting the cluster partition of firms is that estimating models on more (and smaller) firms now exposes sparsity in the coefficient structure: Lasso-regularized models now perform the best, rather than Ridge-regularized ones. This facilitates an



TABLE 6  
Cluster Versus Industry Predictability

Table 6 is a comparison of overall out-of-sample predictability across the two different partitions of the cross section. Each cell reports an incremental  $R^2_{OOS}$  (%) value, calculated as the difference between cluster-partitioned and industry-partitioned model  $R^2_{OOS}$  performances, together with hypothesis test results, per model type and regularization scheme. The models are each estimated based on all available firms, once per slice. The significance tests are based on 1,000 bootstrap samples. \*\*\*, \*\*, and \* denote significance at the 99%, 95%, and 90% levels, respectively.

Model	Ridge	Lasso
Pooled	-0.03	0.02
By-group	0.20***	0.36***
Two-stage	-0.06***	0.17***

interpretation of which characteristics matter for predictability, as we shall see in Section VI.

For both partitions of the cross section, our aggregate predictability results are also consistent with our prior evidence that incorporating heterogeneity in predictive relationships based on economically interpretable groupings of firms can positively impact OOS predictability.

We now compare the predictive OOS performance of the same model specifications when estimated upon different partitions of the cross section of firms. We calculate the incremental  $R^2_{OOS}$  of cluster-estimated over industry-estimated models, repeating for multiple bootstrap samples in order to calculate bootstrap confidence intervals and therefore test whether the values are significantly non-zero. Table 6 reports these incremental  $R^2_{OOS}$  values.<sup>25</sup> The incremental  $R^2_{OOS}$  values for the pooled models are not significantly different from 0, consistent with the fact they do not make use of firm groupings. The incremental  $R^2_{OOS}$  values of the non-pooled models are mostly positive and significantly nonzero, indicating that exploiting cluster membership rather than industry membership results in better OOS predictive performance. The outperformance is highest for the by-group Lasso model, and outperformance remains when both pooled and by-group stages are combined in a two-stage Lasso-regularized model.

Comparison with Existing Results

We first consider how our results relate to Ross (2005)-style theoretical bounds on return predictability. In a survey, Rapach and Zhou (2013) argue that monthly in-sample  $R^2$  values in the neighborhood of 1% or less “can nevertheless signal ‘too much’ return predictability and the existence of market inefficiencies from the standpoint of existing asset-pricing models.” Our  $R^2_{OOS}$  metric is an OOS statistic, and Rapach and Zhou (2013) note that “OOS  $R^2$  statistics will frequently be even lower,” implying that OOS values of around 1% are even more economically notable than in-sample values. Empirically, some variants of our two-stage predictive strategy in Panel A of Table 5 exhibit an OOS performance level of  $R^2_{OOS} > 1.9\%$

<sup>25</sup>Note that these incremental  $R^2_{OOS}$  values are not equal to the differences between  $R^2_{OOS}$  values reported in Panel C of Table 4 and Panel C of Table 5: The conditions we impose to avoid making arbitrary cluster assignments for new firms (described in Section III.C) lead to different samples in parts of the overall data set, and so we must jointly evaluate and bootstrap the models on the test set samples in common to both the industry and cluster partitions so that the estimates are comparable.

when we train on a subset of the largest firms only and exploit cluster groupings. When training on the entire cross section and exploiting cluster groupings, our best (linear) model achieves an  $R^2_{\text{OOS}} = 1.05\%$  in Panel C.

As for comparisons to the empirical state of the art, the closest framework is that of GKX, since we adopt a similar slicing strategy and the same measure of OOS performance ( $R^2_{\text{OOS}}$ ), albeit with the caveat that our samples differ. Table 1 in GKX indicates that the best neural network methods detailed in that article have an  $R^2_{\text{OOS}} = 0.40\%$  when evaluated on their full sample. Procedurewise, our most comparable results to GKX are those where we do not restrict ourselves to clustered firms: In Panel C of Table 4, our best (linear) model exploits industry groupings to reach an OOS performance level of  $R^2_{\text{OOS}} = 0.76\%$ . When we use the clustering procedure, in Panel C of Table 5, our best (linear) model achieves an  $R^2_{\text{OOS}} = 1.05\%$ . The closest method that GKX used to ours is the Pooled ElasticNet, and we earlier found evidence that allowing for heterogeneous predictive relationships across clusters (i.e., using the by-cluster ElasticNet model) improves on its predictive accuracy.

## VI. Variable Importance and Heterogeneity

Prior studies (Gu, Kelly, and Xiu (2020), Freyberger, Neuhierl, and Weber (2020), and DeMiguel et al. (2020)) took various approaches to answer Cochrane's (2011) questions of "Which characteristics really provide independent information about average returns? Which are subsumed by others?" Such studies have so far assumed that the same set of characteristics must matter for all firms in the cross section. Based on our finding that heterogeneity positively impacts predictability, we suggest an additional, novel channel: Characteristics may matter not only as direct predictors of next-period returns, but also in proxying for firms' latent group memberships. Furthermore, different characteristics may enter into different group-specific predictive relationships.

In Section V.A, we interpreted Figure 2 to describe the clusters of firms based on which firm characteristics varied the most from cluster to cluster. Similarly, Figure 2 also describes which characteristics are most important<sup>26</sup> in forming those clusters of firms: firm AGE (and the related measure of whether they recently IPO'ed), SFE, CHPMIA, SECUREDIND, PCHSALE\_PCHINVT, and OPERPROF. Given this partition of the cross section of firms, we analyze which characteristics directly predict firms' next-month returns by using Lasso-regularized models to discard irrelevant predictive characteristics for each cluster.

Table 7 reports the frequency with which firm-level predictive characteristics are selected for each cluster of firms using the By-Cluster Lasso procedure. This variation of selected predictive characteristics by cluster is further evidence of heterogeneity in the form of predictive relationships that apply to different groups of firms. We focus on the By-Cluster Lasso procedure because frequency of selection is an interpretable measure of variable importance, but would like to

<sup>26</sup>We conduct an alternative analysis of characteristic importance for cluster formation based on the dispersion of cluster centroids in the Supplementary Material, which gives similar results.

TABLE 7  
By-Cluster Lasso Selected Predictors

Table 7 reports the frequency of selection (% of slices) of characteristics by cluster, when estimating the by-cluster Lasso model. The model was estimated based on all available firms, once per slice.

Characteristic	Cluster 1	Cluster 2	Cluster 3	Cluster 4
(Intercept)	100	100	100	100
BASPREAD	17	0	33	0
CASHPR	33	17	33	0
CHPMIA	33	0	33	0
DP_MKT	33	17	17	0
SUE	0	0	17	0

TABLE 8  
By-Cluster Lasso Significant Predictors

Table 8 reports the frequency (% of slices) for which each coefficient is significant (at the 99% level) by cluster, when estimating the by-cluster Lasso model. Significance tests are based on 1,000 bootstrap samples.

Characteristic	Cluster 1	Cluster 2	Cluster 3	Cluster 4
(Intercept)	100	100	33	0
BASPREAD	0	0	33	0
CASHPR	33	17	0	0
CHPMIA	17	0	33	0
DP_MKT	17	17	0	0
SUE	0	0	17	0

highlight that Lasso-regularized models have the best OOS predictive performance, so this interpretability does not come at the expense of performance.

Employing a Lasso penalty selects a subset of coefficients by setting all nonselected coefficients' values to 0, as we see in Table 7. We go further and test whether the coefficients are significantly different from 0 by computing bootstrap confidence intervals for each coefficient. Table 8 reports the results of such a bootstrap analysis: It shows the number of coefficients (displayed as a frequency) that are nonzero at a 99% level of significance. By comparing Tables 7 and 8, it is clear that a similar pattern of predictive variable importance emerges as we test for statistical significance: The main difference is that two coefficients (CASHPR and DP\_MKT) are never significantly nonzero for Cluster 3 even though they were selected by the Lasso for a third of our slices.

Table 9 reports the frequency of selection of characteristics by the two-stage Lasso-regularized procedure, our best-performing model on the entire sample of firms when exploiting the cluster partition. Based on this frequency of selection, cash productivity (CASHPR) and adjusted changes in profit margins (CHPMIA) are important predictors, especially for Cluster 1 (more mature firms). Another firm-level characteristic that often predicts for all clusters is ROIC. The market-level D/P ratio (DP\_MKT) is often important for predicting returns for Cluster 1 (more mature firms). The remaining firm-level characteristics are sometimes selected for all clusters and sometimes for individual ones. Note that the intercept terms represent market-level and cluster-level historical returns, so these are also used for prediction throughout. Notably absent from the list of selected characteristics are stock trends (momentum and price reversal), market beta, book-to-market, and earnings-to-price.

TABLE 9  
Two-Stage Lasso (on Clusters) Selected Predictors

Table 9 reports the frequency of selection (% of slices) of characteristics by cluster, when estimating the two-stage Lasso-regularized model. The model was estimated based on all available firms, once per slice.

Characteristic	(Pooled)	Cluster 1	Cluster 2	Cluster 3	Cluster 4
(Intercept)	100	100	100	100	100
BASPREAD	17	17	0	17	0
CASHPR	50	33	17	17	0
CHPMIA	50	33	0	17	0
DP_MKT	17	33	17	17	0
MVE	17	0	0	0	0
ROIC	33	0	0	0	0
SUE	17	0	0	0	0
TB	17	0	0	0	0

It is worth emphasizing the distinction between grouping variables and predictive variables. Characteristics can determine risk premia in two ways: through a direct predictive link to returns, or as a means to group firms with similar predictive relationships together. The latter aspect has not yet been studied in the literature. To illustrate this distinction, consider the AGE firm-level characteristic: Jiang et al. (2005) found empirical evidence that the AGE variable predicts returns (and interpret their findings through the lens of “information uncertainty”), yet this is unsupported by our findings here. Rather, we find that for the cluster containing more mature (and hence older) firms, the firm-level characteristics and market-level variables described above tend to be selected. Therefore, AGE is an important firm-level characteristic insofar as it helps us to identify a grouping of mature firms, rather than a predictor of firm-level returns itself. This distinction may be important, and may also help rationalize the apparent existence of many variables that appear to predict firm-level returns in previous work, but whose importance is of a different nature altogether. Indeed, if only a few variables directly predict next-period returns, this is consistent with the notion of sparsity in the cross section.

## Comparison with Existing Results

We differ from the literature in our findings of what characteristics appear to be most important for firm-level predictability: GKX (see also Freyberger et al. (2020)) find the most informative stock-level predictors fall into three categories. First and most informative of all are price trend variables (e.g., stock momentum, industry momentum, and short-term reversal). The next are liquidity variables (e.g., market value, dollar volume, and bid–ask spread). Finally, return volatility, idiosyncratic volatility, market beta, and beta squared are also among the leading predictors in all models they consider.

We use a different measure of variable predictive importance to GKX (namely, the frequency of selection by sparse models) and find that, in contrast to their results, a very small subset of low-frequency cash and profitability-related coefficients (CHPMIA and CASHPR) and the market-level D/P ratio (DP\_MKT) tend to vary across clusters of firms, whereas a few other variables (BASPREAD, ROIC, MVE, and SUE) are only selected at the level that is common to all firms in the cross section. The good OOS performance of Lasso-regularized models that use only

these selected coefficients confirms their importance.<sup>27</sup> Our analysis in the next section shows that portfolios formed based on this sparse set of signals deliver economically meaningful Sharpe ratios OOS. In general, the parsimony that we uncover echoes Kelly et al. (2019), who found that only a small subset of stock characteristics was responsible for IPCA's empirical performance by better identifying dynamic latent factor loadings. Feng et al. (2020) also seek (and find) parsimony among a high-dimensional set of asset-pricing factors.

It is worth noting that economic agents may use a sparse and parsimonious view of the world, due, for example, to bounded rationality (Gabaix (2014)) or model uncertainty (Guecioueur (2020)), and their demand functions may therefore depend only on a sparse set of characteristics (Nagel (2021)). Interesting topics for future research are linking such sparse demand functions to the sparsity in return predictability that we detect in our setting, and linking investor clienteles to our estimated partition of the cross section of firms.

We also contribute by providing statistical evidence for the importance of our selected predictive variables using the bootstrap, in contrast to the nonstatistical measures used in the prior literature.

## VII. Portfolio Analysis

We now quantify the economic significance of the predicted firm-level expected returns. In doing so, we develop 3 sets of portfolio analyses which confirm that the potential economic gains of our approach are substantial.

In each portfolio analysis, we report the following summary statistics for each portfolio return series: the (annualized) SR, the mean monthly return and its standard deviation, the average portfolio turnover, the maximum drawdown, and the estimated alpha with respect to a 6-factor model, comprising the Fama and French (2015) 5 factors plus the Carhart (1997) momentum factor, and its test statistic (based on Newey–West standard errors computed with 12 lags), together with the  $R^2$  of monthly excess returns with respect to this 6-factor model.

Within each portfolio analysis, we report results using two weighting schemes for each leg (long and short): First, a value weighting using the previous month's market capitalization of each stock in the portfolio, and second, an equal weighting of each stock.

As is standard in the literature (see, e.g., GKX; DeMiguel, Garlappi, and Uppal (2009)), the average monthly portfolio turnover is calculated based on the weight  $w_{it}$  of stock  $i$  in the portfolio at month  $t$  as the average sum of the absolute values of the trades across all holdings,

$$(14) \quad \text{Turnover} = \frac{1}{T} \sum_{t=1}^T \left( \sum_i \left| w_{i,t+1} - \frac{w_{it}(1 + r_{i,t+1})}{1 + \sum_j w_{jt} r_{j,t+1}} \right| \right),$$

<sup>27</sup> As a robustness exercise, Appendix D of the Supplementary Material shows the results of re-estimating the by-cluster lasso model on the smallest 1,000 firms in our sample. A slightly more diverse set of 15 predictive variables is selected. Nevertheless, this remains only a subset of the full set of predictive variables used in our study, indicating that our methodology continues to detect a parsimonious set of characteristics that predict returns OOS.

and the maximum drawdown is calculated based on the cumulative log return  $Y_t$  from inception to month  $t$  as

$$(15) \quad \text{MAXDD} = \max_{0 \leq t_1 \leq t_2 \leq T} (Y_{t_1} - Y_{t_2}).$$

The first portfolio analysis takes an approach commonly found in the literature: Each month, firms are ranked by their expected return predictions, and a long portfolio is formed out of the top 30th percentile and a short portfolio is formed out of the bottom 30th percentile. The results of this exercise can be found in Table 10.

The results of this first portfolio analysis indicate that forming long-short portfolios based on the ranks of predicted monthly returns generated by the by-cluster lasso model (which incorporates sparsity in predictive coefficients together with heterogeneity in predictive relationships) delivers an economically meaningful OOS annualized SR of 0.70 on VW holdings or 0.69 on EW holdings. The monthly alphas (0.23% VW or 0.24% EW) are positive and significant (VW at

TABLE 10  
Long-Short Portfolios Based on Rank Percentiles

In Table 10, portfolios are formed by going long stocks in the top 30th percentile of next-month expected returns, and short stocks within the bottom 30th percentile of next-month expected returns. Each panel denotes a portfolio weighting scheme. Results are split into groups according to the partition of firms used, and models are ranked by out-of-sample (OOS) Sharpe ratio (SR) within each such group. *Note:* SRs are presented annualized. "Monthly Return" and "Std. Dev." denote the mean and standard deviation of the monthly long-short portfolio return, respectively. \*\*\*, \*\*, and \* denote significance at the 99%, 95%, and 90% levels, respectively, of the corresponding FF5 + MOM monthly  $\alpha$  based on the test statistic shown alongside it (in parentheses). All values in the table are calculated OOS, that is, from monthly portfolio returns during the OOS test period of 2010 to 2015, inclusive.

Partition	Model	SR	Monthly Return (%)	Std. Dev. (%)	Turnover (%)	Drawdown (%)	FF5 + MOM $\alpha$ (%)	FF5 + MOM $R^2$ (%)
<i>Panel A. Value-Weighted</i>								
Clusters	By-cluster Lasso	0.70	0.23	1.15	18.54	5.57	0.23* (1.87)	2.35
	By-cluster ElasticNet	0.48	0.25	1.83	19.59	15.32	0.12 (0.5)	10.81
	Two-stage Lasso	0.36	0.13	1.25	19.87	10.62	0.09 (0.55)	2.09
	Two-stage Ridge	0.31	0.20	2.29	18.55	23.74	-0.02 (-0.08)	13.54
	By-cluster Ridge	0.26	0.15	1.96	18.51	19.37	-0.05 (-0.19)	13.23
Industries	Two-stage Ridge	0.32	0.20	2.13	19.86	17.54	0.01 (0.05)	9.66
	By-industry Ridge	0.22	0.12	1.85	20.25	18.18	-0.06 (-0.22)	10.20
	By-industry ElasticNet	0.04	0.03	1.84	20.22	18.59	-0.02 (-0.1)	3.43
	Two-stage Lasso	-0.01	0.00	1.79	22.07	16.51	0.01 (0.05)	2.92
	By-industry Lasso	-0.01	0.00	1.79	22.07	16.51	0.01 (0.05)	2.92
None	Pooled OLS	0.18	0.13	2.43	20.51	26.30	-0.06 (-0.16)	10.96
	Pooled Ridge	0.16	0.12	2.64	18.75	31.49	-0.14 (-0.34)	14.17
	Pooled ElasticNet	0.15	0.12	2.65	18.83	31.49	-0.14 (-0.37)	14.41
	Pooled Lasso	-0.93	-0.47	1.76	17.28	26.93	-0.56 (-1.07)	6.84
<i>Panel B. Equal-Weighted</i>								
Clusters	By-cluster Lasso	0.69	0.25	1.26	54.75	5.77	0.24** (2.01)	4.76
	By-cluster ElasticNet	0.59	0.33	1.91	46.48	15.15	0.19 (0.79)	13.41
	Two-stage Lasso	0.42	0.16	1.29	53.75	9.31	0.12 (0.77)	5.11
	Two-stage Ridge	0.38	0.25	2.26	41.91	23.60	0.01 (0.04)	15.36
	By-cluster Ridge	0.33	0.19	1.99	43.10	19.50	-0.01 (-0.06)	14.88
Industries	Two-stage Ridge	0.47	0.29	2.12	43.26	15.78	0.12 (0.54)	11.85
	By-industry Ridge	0.36	0.20	1.85	48.09	16.87	0.03 (0.13)	12.39
	By-industry ElasticNet	0.18	0.10	1.88	50.05	18.05	0.04 (0.18)	4.91
	By-industry Lasso	-0.05	-0.02	1.74	53.76	17.67	-0.02 (-0.08)	1.79
	Two-stage Lasso	-0.05	-0.02	1.74	53.76	17.67	-0.02 (-0.08)	1.79
None	Pooled OLS	0.39	0.26	2.32	47.98	23.46	0.06 (0.15)	12.48
	Pooled Ridge	0.25	0.19	2.56	38.53	30.08	-0.09 (-0.23)	15.05
	Pooled ElasticNet	0.24	0.18	2.55	38.52	30.08	-0.09 (-0.25)	15.40
	Pooled Lasso	-0.93	-0.46	1.73	46.46	27.74	-0.66 (-1.3)	9.31

the 90% level or EW at the 95% level) with respect to the FF5 + MOM 6-factor model.<sup>28</sup> Furthermore, the  $R^2$  with respect to the 6-factor model is lower for the portfolio returns that are based on sparse Lasso-regularized predictive models than those based on non-sparse Ridge-regularized models; this suggests that strategies that take advantage of sparsity may uncover sources of predictability that are not well explained by the 6-factor model.

While the ranking methodology of the first portfolio analysis is the norm in the empirical asset-pricing literature, an investor aiming to exploit our models' predictions would also take into account the predicted *sign* of the return, which is not explicitly incorporated by portfolios formed on rank percentiles. For example, given a predicted *negative* return, an investor may not necessarily include this stock into her *long* portfolio, irrespective of its rank compared to other stocks. This behavior would be consistent with an investor who prefers more to less.

We therefore repeat a similar analysis but now form long portfolios out of stocks that are predicted to have a positive expected return in the coming month, and form short portfolios out of stocks that are predicted to have a negative expected return in the coming month. This explicitly takes advantage of the availability of a predicted sign for each firm-level expected return. The results of this second portfolio analysis are presented in Table 11.

The results of this second portfolio analysis, in Table 11, show that an investor who forms long-short portfolios based on the predicted sign of next-month expected returns earns a higher (and more economically significant) OOS annualized SR of 1.18 (on VW holdings) or 1.21 (on EW holdings) by using models that incorporate both cluster heterogeneity with sparsity, such as the By-Cluster Lasso or Two-Stage Lasso using clusters (which both produce identical portfolio allocations in this case). Furthermore, the monthly alphas (1.93% VW or 1.96% EW) are positive and significant at the 99% level with respect to the 6-factor model (FF5 + MOM).

Focusing on value weightings in Panel A of each table, a number of other patterns emerge from Tables 10 and 11. First, the portfolio with the highest SR is always based on Lasso regularization and the cluster partition of firms: When forming portfolios based on rank percentiles, the highest SR is earned by following the predictions of the by-cluster Lasso model (SR of 0.70); when forming portfolios based on predicted signs, the highest SR is earned by both the By-Cluster Lasso and Two-Stage Lasso (SR of 1.18) on the cluster partition of firms. Our portfolio analyses are therefore consistent with our findings on aggregate predictability (in Section V.C), where we likewise found that the highest predictive  $R^2_{\text{OOS}}$  values were achieved by Lasso-regularized models on the cluster partition of firms. Second, pooled models that ignore the information given by the cluster or industry groupings of firms never earn the highest SR. Third, the highest  $R^2$  with respect to the FF5 + MOM factor model anywhere in these two tables is 32%, and most values tend to be below 10%, suggesting a pattern of predictability that is distinct from these risk factors and thus hard to explain with the 6-factor model. Finally, comparing EW portfolios in Panel B of each table to the VW portfolios in Panel A, EW portfolios

<sup>28</sup>The usual caveat regarding the “bad-model” problem (Fama (1998)) applies when discussing performance evaluation.



TABLE 11  
Long-Short Portfolios Based on Predicted Signs

As reported in Table 11, portfolios are formed by going long stocks with positive next-month expected returns, and short stocks with negative next-month expected returns. Each panel denotes a portfolio weighting scheme. Results are split into groups according to the partition of firms used, and models are ranked by out-of-sample (OOS) Sharpe ratio (SR) within each such group. *Note:* Two-stage Lasso models produce the same predicted signs as the corresponding by-cluster/industry Lasso models, resulting in identical long-short portfolio compositions, and are therefore omitted from the table for brevity. SRs are presented annualized. "Monthly Return" and "Std. Dev." denote the mean and standard deviation of the monthly long-short portfolio return, respectively. \*\*\*, \*\*, and \* denote significance at the 99%, 95%, and 90% levels, respectively, of the corresponding FF5 + MOM monthly  $\alpha$  based on the test statistic shown alongside it (in parentheses). All values in the table are calculated OOS, that is, from monthly portfolio returns during the OOS test period of 2010 to 2015, inclusive.

Partition	Model	SR	Monthly return (%)	Std. Dev. (%)	Turnover (%)	Drawdown (%)	FF5 + MOM $\alpha$ (%)	FF5 + MOM $R^2$ (%)
<i>Panel A. Value-Weighted</i>								
Clusters	By-cluster Lasso	1.18	1.68	4.91	14.73	26.17	1.93*** (3.77)	9.17
	By-cluster ElasticNet	0.87	1.29	5.12	29.48	26.17	1.58*** (3.03)	5.99
	By-cluster Ridge	0.66	0.95	4.96	21.08	26.17	1.16** (2.46)	6.12
	Two-stage Ridge	0.34	0.50	5.04	27.21	30.64	0.6 (1.18)	8.09
Industries	By-industry Ridge	0.71	1.00	4.83	19.95	27.71	1.11*** (2.77)	8.59
	Two-stage Ridge	0.44	0.64	5.00	27.19	27.71	0.71 (1.58)	7.39
	By-industry Lasso	0.43	1.16	9.23	27.54	50.27	0.2 (0.19)	31.55
	By-industry ElasticNet	-0.13	-0.35	9.25	27.85	128.33	-0.37 (-0.36)	9.56
None	Pooled Lasso	0.92	1.28	4.78	7.00	26.17	1.44*** (2.82)	9.49
	Pooled OLS	0.58	0.59	3.49	129.28	21.26	0.52 (1.16)	12.27
	Pooled ElasticNet	0.46	0.67	4.96	27.13	26.17	0.75* (1.67)	9.29
	Pooled Ridge	0.43	0.61	4.95	27.33	26.17	0.68 (1.51)	9.20
<i>Panel B. Equal-Weighted</i>								
Clusters	By-cluster Lasso	1.21	1.74	4.95	37.02	26.31	1.96*** (3.67)	8.67
	By-cluster ElasticNet	0.86	1.31	5.25	51.98	26.31	1.56*** (2.84)	5.27
	By-cluster Ridge	0.66	0.98	5.08	42.09	26.31	1.13** (2.25)	5.83
	Two-stage Ridge	0.38	0.56	5.18	49.51	28.21	0.63 (1.24)	7.83
Industries	By-industry Ridge	0.75	1.06	4.88	40.71	28.09	1.14*** (2.68)	8.74
	Two-stage Ridge	0.49	0.73	5.14	48.19	28.09	0.73 (1.63)	7.62
	By-industry Lasso	0.45	1.21	9.30	48.81	50.76	0.16 (0.15)	32.18
	By-industry ElasticNet	-0.09	-0.25	9.27	48.51	128.10	-0.33 (-0.32)	10.00
None	Pooled Lasso	0.93	1.31	4.85	27.98	26.31	1.43** (2.6)	9.27
	Pooled OLS	0.65	0.70	3.72	142.85	24.54	0.59 (1.25)	13.97
	Pooled ElasticNet	0.53	0.78	5.04	49.67	26.31	0.82* (1.87)	9.62
	Pooled Ridge	0.49	0.72	5.05	49.84	26.31	0.74* (1.69)	9.48

tend to earn a higher SR, consistent with the findings of GKX. However, this comes at the cost of a slightly higher monthly turnover for the EW portfolios.

Not only do By-Cluster Lasso portfolios have higher SRs than alternatives, but they also have comparatively small drawdowns. The maximum drawdowns experienced for the By-Cluster Lasso strategy based on rank percentiles and predicted signs are 5.77% and 26.31%, respectively, for equal weights (and slightly lower for VW portfolios). Turnover is consistently between 15% and 55% per month. This low turnover, even for EW portfolios which tend to have higher turnover, may be a sign of robustness in the presence of transaction costs, as indicated by Novy-Marx and Velikov (2016).<sup>29</sup> As a frame of reference, the monthly turnovers of the best-performing strategies in GKX, based on neural networks, exceed 110% per month. They attribute this high turnover mostly to the large role of price trend predictors selected by that approach, which lead to a comparatively high portfolio turnover.

<sup>29</sup>As Novy-Marx and Velikov (2016) note, "most anomalies with less than 50% turnover per month generate significant net spreads when designed to mitigate transaction costs; few with higher turnover do."

As we explain in Section VI, we find that, in contrast to their results, a very small subset of low-frequency cash and profitability-related coefficients (CHPMIA and CASHPR) and the market-level D/P ratio (DP\_MKT) tend to vary across clusters of firms, whereas a few other variables (BASPREAD, ROIC, MVE, and SUE) are only selected at the level that is common to all firms in the cross section. Notably absent from our list of selected firm-level characteristics are stock trends (momentum and price reversal), market beta, book-to-market, and earnings-to-price.

Our third and final portfolio analysis considers a hybrid method for forming portfolios, and also serves as a robustness check that using predicted return signs generates economic gains. In this procedure, the top 50th percentile of the predicted-positive-return stocks are used to form the long portfolios, and the bottom 50th percentile of the predicted-negative-return stocks are used to form the short portfolios. The results of applying this hybrid methodology are shown in Table 12, and are qualitatively similar to the previous set of results described in Table 11, with one notable addition: All portfolios that use industry partitions now achieve

TABLE 12  
Long-Short Portfolios Based on Signs and Ranks

As reported in Table 12, portfolios are formed by going long the top 50th percentile of stocks with positive next-month expected returns, and short the bottom 50th percentile of stocks with negative next-month expected returns. Each panel denotes a portfolio weighting scheme. Results are split into groups according to the partition of firms used, and models are ranked by out-of-sample (OOS) Sharpe ratio (SR) within each such group. *Note:* Two-stage lasso models on the industry partition result in identical long-short portfolio compositions to by-industry Lasso models, and are therefore omitted from the table for brevity. SRs are presented annualized. "Monthly Return" and "Std. Dev." Denote the mean and standard deviation of the monthly long-short portfolio return, respectively. \*\*\*, \*\*, and \* denote significance at the 99%, 95%, and 90% levels, respectively, of the corresponding FF5 + MOM monthly  $\alpha$  based on the test statistic shown alongside it (in parentheses). All values in the table are calculated OOS, that is, from monthly portfolio returns during the OOS test period of 2010 to 2015, inclusive.

Partition	Model	SR	Monthly Return (%)	Std. Dev. (%)	Turnover (%)	Drawdown (%)	FF5 + MOM $\alpha$ (%)	FF5 + MOM $R^2$ (%)
<i>Panel A. Value-Weighted</i>								
Clusters	By-cluster Lasso	1.19	1.62	4.71	10.49	27.40	1.89*** (4.07)	8.55
	Two-stage Lasso	1.15	1.58	4.74	10.17	27.40	1.85*** (3.92)	8.76
	By-cluster ElasticNet	0.88	1.33	5.21	18.61	28.46	1.64*** (3.68)	5.95
	By-cluster Ridge	0.84	1.27	5.24	14.14	28.45	1.47** (2.63)	7.73
	Two-stage Ridge	0.84	1.25	5.15	15.46	29.68	1.31** (2.4)	6.98
Industries	By-industry Lasso	0.81	1.37	5.86	25.75	29.31	1.32*** (2.76)	12.03
	Two-stage Ridge	0.77	1.16	5.24	19.78	29.21	1.22** (2.19)	8.26
	By-industry ElasticNet	0.62	1.02	5.68	21.13	29.50	0.77** (2.11)	10.96
	By-industry Ridge	0.60	0.94	5.42	16.80	28.41	0.9** (2.06)	9.36
None	Pooled Lasso	0.82	1.16	4.89	10.68	26.17	1.32** (2.52)	9.53
	Pooled ElasticNet	0.76	1.16	5.30	17.12	29.56	1.18* (1.96)	7.13
	Pooled Ridge	0.76	1.16	5.29	17.08	29.56	1.17* (1.96)	7.18
	Pooled OLS	0.72	0.95	4.58	115.72	27.32	0.76 (1.13)	19.49
<i>Panel B. Equal-Weighted</i>								
Clusters	By-cluster Lasso	1.21	1.66	4.75	30.00	27.50	1.9*** (3.95)	8.12
	Two-stage Lasso	1.18	1.62	4.76	29.22	27.50	1.86*** (3.9)	8.42
	Two-stage Ridge	0.89	1.31	5.11	29.63	28.57	1.35** (2.44)	6.70
	By-cluster ElasticNet	0.87	1.34	5.32	35.84	28.02	1.64*** (3.44)	5.79
	By-cluster Ridge	0.81	1.26	5.34	29.42	27.79	1.42** (2.34)	7.70
Industries	Two-stage Ridge	0.85	1.28	5.19	31.67	28.82	1.31** (2.45)	8.42
	By-industry Lasso	0.82	1.43	6.01	44.07	29.53	1.34*** (2.74)	12.90
	By-industry ElasticNet	0.67	1.10	5.67	36.25	29.37	0.81** (2.34)	12.09
	By-industry Ridge	0.65	1.02	5.40	31.06	28.51	0.96** (2.23)	10.11
None	Pooled OLS	0.90	1.31	5.02	125.54	24.78	1.08 (1.57)	18.94
	Pooled Lasso	0.85	1.21	4.89	28.12	26.31	1.32** (2.42)	9.18
	Pooled ElasticNet	0.84	1.26	5.15	28.61	28.12	1.26** (2.17)	6.81
	Pooled Ridge	0.84	1.25	5.14	28.59	28.12	1.25** (2.17)	6.86

monthly alphas that are significant at the 95% level or higher, with corresponding increases in the economic significance of their OOS SRs. For example, the By-Industry Lasso long-short portfolio delivers an annualized OOS SR of 0.81 (VW holdings) and 0.82 (EW holdings), and the Two-Stage Ridge long-short portfolio delivers an annualized OOS SR of 0.77 (VW holdings) or 0.85 (EW holdings).

We would like to end this analysis by highlighting three ways in which these findings have important practical implications for investors.

First, as evidenced by the existence of sector-specific fund mandates and sector ETFs, investors may choose to form portfolios restricted to an individual sector of the economy. The literature reflects also scholarly attention: For example, Huang, O'Hara, and Zhong (2021) find that investors may hedge at the industry level (using ETFs), Menzly and Ozbas (2010) find that informed investors may specialize in industries, and Peng and Xiong (2006) demonstrate theoretically why investors may choose to process industry-specific information instead of firm-specific information. The ability of our predictive models to exploit sector-specific information given by industry partitions while delivering economically significant SRs suggests that these predictive models may be used as concrete guidance to target exposures to firms in specific segments of the economy.

Second, the superior performance of portfolios formed based on Lasso-regularized predictive models deserves special attention. Since the sparse Lasso-regularized models in our study select only a handful of characteristics as having predictive worth, the corresponding portfolios are similarly formed based on only a few predictive signals that have been selected from a wider set. Furthermore, the signal combination step is linear, thanks to the functional form of the predictive models in our study. This is noteworthy because Novy-Marx (2016) singles out the related case of selecting and combining portfolio signals as the setting for his study on potential overfitting bias, where signals may appear to perform well in-sample before degenerating OOS. In our study, we took special care in our empirical framework to avoid over-fitting (including splits into training, validation, test sets, care in hyperparameter tuning, avoidance of look-ahead bias, and so on). The true OOS performance of the portfolios that we form is economically significant while avoiding the sorts of concerns that Novy-Marx (2016) highlights.

Finally, the significant performance attained by portfolio trading strategies that are based on cross-sectional predictions of firm-level monthly returns suggests that detecting cross-sectional variation in the conditional mean function of firm-level returns delivers a profitable trading strategy for investors to follow. This is consistent with the findings of Conrad and Kaul (1998) and DeMiguel, Nogales, and Uppal (2014) that cross-sectional variation in the mean returns of the individual stocks held in a portfolio is an important determinant of the portfolio's profitability.

## VIII. Conclusion

We develop an approach that combines the estimation of linear models of expected excess returns with the assignment of (possibly) latent group membership to firms, both based on observable characteristics. Our procedure incorporates an economically motivated notion of heterogeneity in the cross section that favorably impacts our ability to predict next-period returns. We use a nonparametric

bootstrap approach to underpin our findings with statistical evidence. To select which characteristics matter, we use sparse linear models to discard irrelevant predictive variables and test which characteristics were most frequently selected by our sparse models.

We find models that incorporate predictive heterogeneity in the cross section of firms outperform models that ignore this information. Extracting useful signals from an avalanche of data remains a challenging endeavor, and our findings speak to the literature on understanding the factor “zoo” (Cochrane (2011)).

Our model-building approach contributes by providing evidence on what structure the functional form of return prediction models estimated using ML techniques should take. When left unrestricted by the econometrician, there is no guarantee that the outcome of estimating any model has an economic interpretation, so we facilitate interpretability by specifying an economically grounded structure that carries over to practical aspects of portfolio choice.

Furthermore, using  $k$ -means clustering to partition firms by characteristics (rather than relying on industry membership) not only results in improved OOS predictive performance, but also highlights that only a few characteristics matter for directly predicting next-period returns. On the other hand, different characteristics are important for grouping together firms with common predictive relationships, suggesting the novel implication that characteristics play two roles in determining next-period returns: as direct predictors (the subject of much prior work) and for inferring firm groupings.

Our uncovering of sparsity in predictive variables mirrors the recent “taming” of the factor zoo by Feng et al. (2020), who also seek parsimony in a high-dimensional setting. Arguments against the existence of sparsity in economics (Giannone, Lenza, and Primiceri (2021)) typically do not acknowledge the possibility of group-varying predictive relationships, as we do, and our detection of sparsity together with heterogeneity in the cross section of returns is therefore relevant for other economic prediction problems. While the outmoded practice of manually selecting important variables has been superseded by ML approaches, as Nagel (2021) highlights, our findings suggest that sparsity does have a role to play in cross-sectional return prediction when it is the outcome of an estimated model.

Future work could study various avenues of introducing explicit nonlinearities into the predictive step of our procedure, for example, by projecting characteristics onto a suitable basis, potentially in conjunction with a model selection step. Furthermore, given the informativeness of our stable cluster-based partition of the cross section for detecting conditional risk premia, future work could employ clustering at different frequencies to extract alternative signals relating to the evolution of the population composition of listed firms, as in Brown and Kapadia (2007).

## Supplementary Material

To view supplementary material for this article, please visit <http://doi.org/10.1017/S0022109022001028>.

## References

- Ando, T., and J. Bai. "Clustering Huge Number of Financial Time Series: A Panel Data Approach with High-Dimensional Predictors and Factor Structures." *Journal of the American Statistical Association*, 112 (2017), 1182–1198.
- Asness, C. S.; R. B. Porter; and R. L. Stevens. "Predicting Stock Returns Using Industry-Relative Firm Characteristics." *Available at SSRN*, 213872 (2000).
- Balasubramaniam, V.; J. Y. Campbell; T. Ramadorai; and B. Ranish. "Who Owns What? A Factor Model for Direct Stockholding." *Journal of Finance*, forthcoming (2023).
- Barrot, J.-N., and J. Sauvagnat. "Input Specificity and the Propagation of Idiosyncratic Shocks in Production Networks." *Quarterly Journal of Economics*, 131 (2016), 1543–1592.
- Belloni, A.; D. Chen; V. Chernozhukov; and C. Hansen. "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain." *Econometrica*, 80 (2012), 2369–2429.
- Belloni, A.; V. Chernozhukov; and C. Hansen. "Inference on Treatment Effects After Selection Among High-Dimensional Controls." *Review of Economic Studies*, 81 (2014), 608–650.
- Bonhomme, S., and E. Manresa. "Grouped Patterns of Heterogeneity in Panel Data." *Econometrica*, 83 (2015), 1147–1184.
- Brown, G., and N. Kapadia. "Firm-Specific Risk and Equity Market Development." *Journal of Financial Economics*, 84 (2007), 358–388.
- Cameron, A. C., and P. K. Trivedi. *Microeconometrics: Methods and Applications*. New York, NY: Cambridge University Press (2005).
- Campbell, J. Y., and S. B. Thompson. "Predicting Excess Stock Returns Out of Sample: Can Anything Beat the Historical Average?" *Review of Financial Studies*, 21 (2008), 1509–1531.
- Carhart, M. M. "On Persistence in Mutual Fund Performance." *Journal of Finance*, 52 (1997), 57–82.
- Chernick, M. R. *Bootstrap Methods: A Guide for Practitioners and Researchers*, 2nd ed. Hoboken, NJ: John Wiley & Sons (2007).
- Chernozhukov, V.; D. Chetverikov; M. Demirer; E. Duflo; C. Hansen; and W. Newey. "Double/Debiased/Neyman Machine Learning of Treatment Effects." *American Economic Review Papers and Proceedings*, 107 (2017), 261–265.
- Chetty, R.; A. Looney; and K. Kroft. "Salience and Taxation: Theory and Evidence." *American Economic Review*, 99 (2009), 1145–1477.
- Cochrane, J. H. "Presidential Address: Discount Rates." *Journal of Finance*, 66 (2011), 1047–1108.
- Cohen, L., and A. Frazzini. "Economic Links and Predictable Returns." *Journal of Finance*, 63 (2008), 1977–2011.
- Conrad, J., and G. Kaul. "An Anatomy of Trading Strategies." *Review of Financial Studies*, 11 (1998), 489–519.
- Daniel, K.; L. Mota; S. Rottke; and T. Santos. "The Cross-Section of Risk and Returns." *Review of Financial Studies*, 33 (2020), 1927–1979.
- DeMiguel, V.; L. Garlappi; and R. Uppal. "Optimal Versus Naive Diversification: How Inefficient is the 1/N Portfolio Strategy?" *Review of Financial Studies*, 22 (2009), 1915–1953.
- DeMiguel, V.; A. Martin-Utrera; F. J. Nogales; and R. Uppal. "A Transaction-Cost Perspective on the Multitude of Firm Characteristics." *Review of Financial Studies*, 33 (2020), 2180–2222.
- DeMiguel, V.; F. J. Nogales; and R. Uppal. "Stock Return Serial Dependence and Out-of-Sample Portfolio Performance." *Review of Financial Studies*, 27 (2014), 1031–1073.
- Diebold, F., and R. Mariano. "Comparing Predictive Accuracy." *Journal of Business and Economic Statistics*, 13 (1995), 253–263.
- Diebold, F. X., and M. Shin. "Machine Learning for Regularized Survey Forecast Combination: Partially-Egalitarian Lasso and Its Derivatives." *International Journal of Forecasting*, 35 (2019), 1679–1691.
- Ding, C., and X. He. "K-Means Clustering via Principal Component Analysis." In *Proceedings of the Twenty-First International Conference on Machine Learning*. New York, NY: Association for Computing Machinery (2004), 29.
- Dorn, D., and G. Huberman. "Preferred Risk Habitat of Individual Investors." *Journal of Financial Economics*, 97 (2010), 155–173.
- Fama, E. F. "Market Efficiency, Long-Term Returns, and Behavioral Finance." *Journal of Financial Economics*, 49 (1998), 283–306.
- Fama, E. F., and K. R. French. "A Five-Factor Asset Pricing Model." *Journal of Financial Economics*, 116 (2015), 1–22.
- Farmer, L.; L. Schmidt; and A. Timmermann. "Pockets of Predictability." *Available at SSRN*, 3152386 (2019).
- Feng, G.; S. Giglio; and D. Xiu. "Taming the Factor Zoo: A Test of New Factors." *Journal of Finance*, 75 (2020), 1327–1370.

- Fisher, J. D.; D. W. Puelz; and C. M. Carvalho. "Monotonic Effects of Characteristics on Returns." *Annals of Applied Statistics*, 14 (2020), 1622–1650.
- Freyberger, J.; A. Neuhierl; and M. Weber. "Dissecting Characteristics Nonparametrically." *Review of Financial Studies*, 33 (2020), 2326–2377.
- Fuster, A.; P. Goldsmith-Pinkham; T. Ramadorai; and A. Walther. "Predictably Unequal? The Effects of Machine Learning on Credit Markets." *Journal of Finance*, 77 (2022), 5–47.
- Gabaix, X. "A Sparsity-Based Model of Bounded Rationality." *Quarterly Journal of Economics*, 129 (2014), 1661–1710.
- Gabaix, X. "Behavioral Inattention." In *Handbook of Behavioral Economics: Applications and Foundations*, Vol. 2. Amsterdam, Netherlands: Elsevier (2019), 261–343.
- Gabaix, X. "A Behavioral New Keynesian Model." *American Economic Review*, 110 (2020), 2271–2327.
- Giannone, D.; M. Lenza; and G. E. Primiceri. "Economic Predictions with Big Data: The Illusion of Sparsity." *Econometrica*, 89 (2021), 2409–2437.
- Green, J.; J. R. Hand; and X. F. Zhang. "The Characteristics That Provide Independent Information About Average US Monthly Stock Returns." *Review of Financial Studies*, 30 (2017), 4389–4436.
- Grishchenko, O. V., and M. Rossi. "The Role of Heterogeneity in Asset Pricing: The Effect of a Clustering Approach." *Journal of Business & Economic Statistics*, 30 (2012), 297–311.
- Gu, S.; B. Kelly; and D. Xiu. "Empirical Asset Pricing via Machine Learning." *Review of Financial Studies*, 33 (2020), 2223–2273.
- Gu, S.; B. Kelly; and D. Xiu. "Autoencoder Asset Pricing Models." *Journal of Econometrics*, 222 (2021), 429–450.
- Guecicœur, A. "How Do Investors Learn as Data Becomes Bigger? Evidence from a FinTech Platform." Available at SSRN, 3708476 (2020).
- Han, Y.; A. He; D. Rapach; and G. Zhou. "Expected Stock Returns and Firm Characteristics: E-LASSO, Assessment, and Implications." Available at SSRN, 3185335 (2021).
- Hanna, R.; S. Mullainathan; and J. Schwartzstein. "Learning Through Noticing: Theory and Evidence from a Field Experiment." *Quarterly Journal of Economics*, 129 (2014), 1311–1353.
- Hastie, T.; R. Tibshirani; and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer Science & Business Media (2009).
- Hoberg, G., and G. Phillips. "Text-Based Network Industries and Endogenous Product Differentiation." *Journal of Political Economy*, 124 (2016), 1423–1465.
- Hou, K. "Industry Information Diffusion and the Lead-Lag Effect in Stock Returns." *Review of Financial Studies*, 20 (2007), 1113–1138.
- Hou, K., and D. T. Robinson. "Industry Concentration and Average Stock Returns." *Journal of Finance*, 61 (2006), 1927–1956.
- Huang, S.; M. O'Hara; and Z. Zhong. "Innovation and Informed Trading: Evidence from Industry ETFs." *Review of Financial Studies*, 34 (2021), 1280–1316.
- Jiang, G.; C. M. Lee; and Y. Zhang. "Information Uncertainty and Expected Returns." *Review of Accounting Studies*, 10 (2005), 185–221.
- Kapetanios, G. "A Bootstrap Procedure for Panel Data Sets with Many Cross-Sectional Units." *Econometrics Journal*, 11 (2008), 377–395.
- Karolyi, G. A., and S. Van Nieuwerburgh. "New Methods for the Cross-Section of Returns." *Review of Financial Studies*, 33 (2020), 1879–1890.
- Kelly, B. T.; S. Pruitt; and Y. Su. "Characteristics are Covariances: A Unified Model of Risk and Return." *Journal of Financial Economics*, 134 (2019), 501–524.
- Kojien, R. S. J., and M. Yogo. "A Demand System Approach to Asset Pricing." *Journal of Political Economy*, 127 (2019), 1475–1515.
- Lee, J. D.; D. L. Sun; Y. Sun; and J. E. Taylor. "Exact Post-Selection Inference, with Application to the Lasso." *Annals of Statistics*, 44 (2016), 907–927.
- Lewellen, J. "The Time-Series Relations Among Expected Return, Risk, and Book-to-Market." *Journal of Financial Economics*, 54 (1999), 5–43.
- Lewellen, J. "The Cross-Section of Expected Stock Returns." *Critical Finance Review*, 4 (2015), 1–44.
- Lien, D., and Q. H. Vuong. "Selecting the Best Linear Regression Model: A Classical Approach." Working Paper No. 606, California Institute of Technology Social Science (1986).
- Lustig, H.; S. Van Nieuwerburgh; and A. Verdelhan. "The Wealth-Consumption Ratio." *Review of Asset Pricing Studies*, 3 (2013), 38–94.
- Menzly, L., and O. Ozbas. "Market Segmentation and Cross-Predictability of Returns." *Journal of Finance*, 65 (2010), 1555–1580.
- Menzly, L.; T. Santos; and P. Veronesi. "Understanding Predictability." *Journal of Political Economy*, 112 (2004), 1–47.



- Merton, R. C. "An Intertemporal Capital Asset Pricing Model." *Econometrica: Journal of the Econometric Society*, 41 (1973), 867–887.
- Moskowitz, T. J., and M. Grinblatt. "Do Industries Explain Momentum?" *Journal of Finance*, 54 (1999), 1249–1290.
- Nagel, S. *Machine Learning in Asset Pricing*. Princeton, NJ: Princeton University Press (2021).
- Novy-Marx, R., "Testing Strategies Based on Multiple Signals." Working Paper, University of Rochester (2016).
- Novy-Marx, R., and M. Velikov. "A Taxonomy of Anomalies and Their Trading Costs." *Review of Financial Studies*, 29 (2016), 104–147.
- Patton, A. J., and B. Weller. "Risk Price Variation: The Missing Half of Empirical Asset Pricing." *Review of Financial Studies*, 35 (2022), 5127–5184.
- Peng, L., and W. Xiong. "Investor Attention, Overconfidence and Category Learning." *Journal of Financial Economics*, 80 (2006), 563–602.
- Rapach, D., and G. Zhou. "Forecasting Stock Returns." In *Handbook of Economic Forecasting*, Vol. 2. Amsterdam, Netherlands: Elsevier (2013), 328–383.
- Rapach, D. E.; J. K. Strauss; J. Tu; and G. Zhou. "Industry Return Predictability: A Machine Learning Approach." *Journal of Financial Data Science*, 1 (2019), 9–28.
- Rapach, D. E.; J. K. Strauss; and G. Zhou. "Out-of-Sample Equity Premium Prediction: Combination Forecasts and Links to the Real Economy." *Review of Financial Studies*, 23 (2010), 821–862.
- Rapach, D. E., and G. Zhou. "Chapter 1: Time-Series and Cross-Sectional Stock Return Forecasting: New Machine Learning Methods." In *Machine Learning for Asset Management*. Hoboken, NJ: John Wiley & Sons (2020), 1–33.
- Reis, R. "Inattentive Consumers." *Journal of Monetary Economics*, 53 (2006), 1761–1800.
- Ross, S. A. "The Arbitrage Theory of Capital Asset Pricing." *Journal of Economic Theory*, 13 (1976), 341–360.
- Ross, S. A. *Neoclassical Finance*. Princeton, NJ: Princeton University Press (2005).
- Rousseeuw, P. J. "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis." *Journal of Computational and Applied Mathematics*, 20 (1987), 53–65.
- Santos, T., and P. Veronesi. "Labor Income and Predictable Stock Returns." *Review of Financial Studies*, 19 (2006), 1–44.
- Sims, C. A. "Implications of Rational Inattention." *Journal of Monetary Economics*, 50 (2003), 665–690.
- Tibshirani, R. J.; J. Taylor; R. Lockhart; and R. Tibshirani. "Exact Post-Selection Inference for Sequential Regression Procedures." *Journal of the American Statistical Association*, 111 (2016), 600–620.
- Timmermann, A. "Forecasting Methods in Finance." *Annual Review of Financial Economics*, 10 (2018), 449–479.
- Welch, I., and A. Goyal. "A Comprehensive Look at the Empirical Performance of Equity Premium Prediction." *Review of Financial Studies*, 21 (2007), 1455–1508.
- Zou, H., and T. Hastie. "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society: Series B (Methodological)*, 67 (2005), 301–320.