Uncovering sparsity and heterogeneity in firm-level return predictability using machine learning

Theodoros Evgeniou, Ahmed Guecioueur, Rodolfo Prieto*

January 10, 2021

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 upon 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. Our results uncover sparsity together with firms' heterogeneity based on their characteristics, improving both predictability and interpretability. 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.

Keywords: Characteristics, Sparsity, Heterogeneity, Industries, Lasso, Clustering, Return Prediction, Big Data.

JEL classification: G1, G17, C55, C58

*INSEAD, Bd de Constance, 77300 Fontainebleau, France, e-mail: theodoros.evgeniou@insead.edu, ahmed.guecioueur@insead.edu, and rodolfo.prieto@insead.edu. We would like to thank Panos Mavrokonstantis for excellent research assistance while he was a Senior Research Scientist at INSEAD. We would also like to thank participants at the 2021 AFA PhD Poster Session, the 2020 European Winter Meetings of the Econometric Society, the 22nd INFER Annual Conference, the 9th INSEAD-Wharton Doctoral Consortium and the INSEAD Accounting & Finance PhD seminar series, as well as Victor DeMiguel, Joël Peress, Marcel Rindisbacher, Raman Uppal and Jinyuan Zhang for their helpful comments. A previous version of this paper was circulated under the title "Modeling heterogeneity in firm-level return predictability with machine learning".

1 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 candidate as it is a low signal-to-noise ratio problem with many potential predictive signals. ML methods can help in the search for the combination of conditioning variables and cross-sectional factors that best describes the returns of individual assets.

Gu et al. (2020b) (henceforth **GKX**) 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 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 approaches 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.[1]

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, we focus on how group membership may determine firm-level risk premia.[2] 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 (see Patton and Weller (2019)), a regularity that may be connected to the fact that different firm characteristics influence investors differently, signaling for example preferred risk habitats (see Dorn and Huberman (2010)). We diverge from the approach taken by existing studies of firm-level

[1] One route is to embed ML methods in equilibrium models, as Fuster et al. (2020) do; another route is to incorporate economic structure into the ML methods themselves, as we do.

[2] We use the terms "expected return" and "risk premium" interchangeably.

return predictability that fit a single model to all firms (we label them "pooled" models), by estimating "by-group" models and testing their incremental out-of-sample performance.

As in GKX, we use firm characteristics and market-wide variables as predictors and evaluate all our models out-of-sample. We differ by using characteristics to infer group membership, and by restricting ourselves to linear models, for a number of reasons. Firstly, linear ML models and their interpretation in terms of regularization and shrinkage are now well understood.[3] Secondly, linear models make testing via bootstrapping computationally easier, in contrast to much costlier nonlinear methods (e.g. neural networks). The bootstrap is a general-purpose method for assessing statistical significance, which we therefore apply to both out-of-sample performance measures and to parameter estimates. Prior studies have used a variety of non-statistical measures for assessing predictive variable importance: we contribute by assessing the statistical significance of Lasso-selected variables.

We next highlight results and contrast two broad grouping criteria: first we group firms according to their industry classifications, measured by SIC codes, and second, we use a standard unsupervised learning technique, k-means clustering, to infer group membership. Clustering groups similar firms together using a standard distance metric, and we find that the number and interpretation of discernible clusters is stable.

Our first result on firm-level return predictability shows that exploiting an economic partition of the cross-section of firms can positively impact out-of-sample predictability. The incremental out-of-sample predictability of by-group over pooled models with the same regularization scheme is positive across the board. We base our inference on a statistical test using confidence intervals computed from bootstrap samples.

We then assess out-of-sample predictive performance on the full cross-section of firms, following the prior literature on firm-level return predictability. We develop flexible two-stage models that capture group-membership predictive relationships, and show that for both grouping criteria these models are the best performing: a Ridge-regularized two-stage model

[3]We use linear models estimated using Lasso, ElasticNet and Ridge penalties. See recent applications by Feng et al. (2020), DeMiguel et al. (2020), Freyberger et al. (2020) and Kozak et al. (2020), among others.

for industry membership and a Lasso-regularized two-stage model for cluster membership. While our sample of firms and time period studied are not the same as for GKX, our out-of-sample R^2s are higher than the best ones attained in that study.

Based on our finding that heterogeneity matters for predictability, we suggest an additional channel to answer Cochrane (2011)'s questions of "Which characteristics really provide independent information about average returns? Which are subsumed by others?": characteristics may matter not only as predictors of next-period returns, but also in proxying for firms' latent group memberships. Furthermore, different (and potentially sparse) sets of characteristics may matter for different groups of firms. This is in contrast to prior studies (Gu et al., 2020b; Freyberger et al., 2020; DeMiguel et al., 2020) which use various approaches assuming that the same set of characteristics matters for all firms in the cross-section.

The distinction between grouping variables and predictive variables is one of the key insights of our approach. Consider the *age* firm-level characteristic: Jiang et al. (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.

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 seven firm-level characteristics and one market-level variable in total, and discards the rest as having no predictive worth. Low-frequency cash and profitability variables matter the most. These results are consistent with the notion that predictive variables are sparse when conditioning on firm groupings, as well as the recent search for parsimony in the asset pricing factors literature (Feng et al., 2020) – a parsimony that is in stark contrast to the existing literature on return predictability. 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.

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 et al. (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 out-of-sample 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 relevant. The closest empirical setup to ours is that of GKX. Han et al. (2020) 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 et al. (2020), which both use spline-based regressions but differ in what procedures

they use to select characteristics: the former uses the Group Lasso, while 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-based regularization for that (shared) goal. We share some objectives with DeMiguel et al. (2020), who take a portfolio choice approach in order to incorporate the effects of transaction costs when determining characteristic importance; we proceed from the standpoint of predictability, rather than portfolio choice.

Green et al. (2017) (henceforth **GHZ**) attempt to discern which firm characteristics provide independent information about monthly stock returns by employing Fama-MacBeth-style contemporaneous regressions. Similarly, Kelly et al. (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.[4] 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 (2019), who study asset-level heterogeneity and segmentation, building upon the clustering techniques of Bonhomme and Manresa (2015). 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 et al., 2020a). 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.

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

[4]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, while Kelly et al. (2019) use characteristics to infer contemporaneous factor loadings.

cross-section between firms' industry memberships and financial returns. Daniel et al. (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 et al. (2000) and Hou (2007).

The remainder of the paper is organized as follows. We present our theoretical background in Section 2 and our empirical framework in Section 3. We explain our data construction procedure in Section 4. In Section 5 we apply our methodology to the data using different grouping specifications and Section 6 treats variable importance and heterogeneity. Section 7 concludes. The Appendix contains all tables. An Internet Appendix provides a review of methods used in the paper, as well as further data details.

2 Background

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

2.1 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,

$$r_{i,t+1} = E_t(r_{i,t+1}) + \epsilon_{i,t+1} = h(c_{it}) + \epsilon_{i,t+1}, \tag{1}$$

where stocks are indexed as $i = 1, \ldots, N_t$, months by $t = 1, \ldots, T$, and c_{it} is an M-dimensional vector of predictors, with an entry in c_{it} corresponding to either a firm *characteristic* or a market-wide 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,

$$h(c_{it}) = \theta' c_{it}, \tag{2}$$

where $\theta \in \mathbb{R}^M$ 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 market-wide factors f_t.[5]

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 information diffuses gradually in financial markets, impacting return predictability along vertical customer-supplier relationships across industries. More recently, Patton and Weller (2019) find evidence of segmentation across their choice of test assets, benchmark factor models and time periods by studying risk premia of the form $E_t(r_{i,t+1}) = \alpha_t I_i + \beta'_i(f_t + \Phi_t I_i)$, where I_i denotes a K-dimensional vector of indicator variables defining firm i's group membership, with $j \in \{1, \ldots, K\}$, and the matrix Φ_t holds the deviation of risk premia for each group j from

[5]The literature on risk premia representation as a function of conditioning variables is vast. For example, Menzly et al. (2004)'s CCAPM features both time varying risk preferences and expectations of dividend growth. Santos and Veronesi (2006)'s CAPM conditioning variables depend on the shares of the firms' dividends and wages over consumption. More recently, Koijen and Yogo (2019) and Kelly et al. (2019) model expected returns and factor loadings explicitly on the asset's own characteristics.

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 (2) so that each firm i also belongs to a single group j (for which the scalar indicator variable $\mathbb{1}_{i\in j}$ takes the value one), and a further set of group-specific coefficients,

$$h_j(c_{it}) = (\alpha_0 + \mathbb{1}_{i\in j}\alpha_j) + (\theta_0 + \mathbb{1}_{i\in j}\theta_j)'c_{it}. \tag{3}$$

In Section 3.1, we explain how we estimate a version of this model.

2.2 Machine learning

With a high-dimensional characteristics vector c_{it}, GKX found that ML methods outperformed traditional ones for the problem 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 θ_0 and θ_j and these are estimated based upon different samples of firms (as we will explain in the methodology section below). We employ ML estimation procedures that penalize/regularize (functions of) the norms of the coefficient vectors θ_0 and θ_j. The choice of coefficient vector norms to be regularized determines whether a Lasso-based, Ridge-based, or ElasticNet-based model is estimated, for the same functional form.[6] We are particularly interested in Lasso-based models, since the Lasso penalization/regularization procedure encourages zero entries in the coefficient vectors. For each ML estimation procedure, the extent of penalization/regularization applied to some coefficient vector θ is determined by the value of the corresponding hyperparameter(s) λ.

[6]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 Internet Appendix for further details of the precise regularization schemes employed.

We evaluate all our models out-of-sample: accordingly, we adopt a training, validation and testing procedure intended to avoid overfitting.

3 Methodology

We present the predictive models in Section 3.1, and develop an empirical framework to benchmark our models by their out-of-sample predictive accuracy and test their statistical significance in Section 3.2. Finally, Section 3.3 describes two grouping procedures: industry membership and k-means clustering.

3.1 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 non-linear models, but each model is estimated on all available firms.[7] In this paper, 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 c_{it} using the same set of estimated coefficients: an intercept α_0 and a $M \times 1$ vector of pooled coefficients θ_0. We write it first in scalar and then in vector form, as follows:

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

$$= \alpha_0 + \theta_0' c_{it}. \qquad \text{(pooled model)} \tag{5}$$

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).

[7] Likewise, Han et al. (2020) 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.

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

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

$$= \alpha_j + \theta_j' c_{it}. \tag{7}$$

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 predictions 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 $\mathbb{1}_{i \in j}$ to denote firm i being a member of group j, we combine these K models as follows:

$$r_{i,t+1} = \sum_{j=1}^{K} \mathbb{1}_{i \in j} (\alpha_j + \theta_j' c_{it}). \qquad \text{(by-group model)} \tag{8}$$

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

Finally, building on our hypothesis that the conditional mean function varies across (groups of) firms in the cross-section, which requires a more flexible specification than a pooled (linear) model to measure, we build a hybrid model that combines pooled models with by-group heterogeneity to form forecasts. This hybrid model is used to compare overall performance and interpret the importance of characteristics.

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:

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

$$= \alpha_0 + \theta_0' c_{it}. \qquad \text{(stage 1, pooled)} \tag{10}$$

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 prediction residuals are now the inputs to the second stage. The second stage estimates one model for each of the K (non-overlapping) groups of firms, each indexed j:

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

$$= \alpha_j + \theta_j' c_{it}. \qquad \text{(stage 2, by-group)} \tag{12}$$

Predictions for firms' excess returns are thus the sums of the forecasts from each stage.[8] The first stage of the model should estimate predictive relationships that are common to all firms in the cross-section, while the second stage should detect any residual predictive relationships that are specific to groups of firms. The two-stage model is linear,

$$r_{i,t+1} = \overbrace{\alpha_0 + \theta_0' c_{it}}^{\text{pooled stage}} + \overbrace{\sum_{j=1}^{K} \mathbb{1}_{i \in j} (\alpha_j + \theta_j' c_{it})}^{\text{by-group stage}}, \qquad \text{(two-stage model)} \tag{13}$$

and we evaluate variants that have been regularized using Lasso and Ridge-based penalties. As in the by-group model, this requires multiple hyperparameters:[9] in this case, $K+1$ hyperparameters $\lambda_0, \lambda_1, \ldots, \lambda_K$ to control the degree of regularization of the $K+1$ coefficient vectors $\theta_0, \theta_1, \ldots, \theta_K$. Note that if pooled models fail to detect some group-specific predictive

[8] 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.

[9] 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.

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 out-of-sample.

Finally, note that each of the individual pooled or group-specific models in our study is a linear function of characteristic inputs 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 indicator variables $\mathbb{1}_{i \in j}$ are fixed/pre-specified: this is the case when industry membership is used to define the values of these indicator variables. Nonlinearities can arise only when the $\mathbb{1}_{i \in j}$ are functions of the characteristics: this is the case when k-means clustering is used to infer the partition of firms.

3.2 Estimation, performance and tests

Following GKX, we split our database into multiple temporal *slices*, with the training set growing by one year with each slice and the subsequent validation and test sets shifting forward by one year each and maintaining a constant size. There are 6 slices in total. The validation set is always 6 years in length. 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.

[Insert Table 1 about here]

The periods covered by each slice's training, validation and test sets are detailed in Table 1. Since we have a shorter overall sample than GKX do, we have shortened each of the three sets in comparison. The purpose of the validation set (as distinct from the training set) is to allow us to tune hyperparameters of our predictive models; specifically, the (multiple) λ parameters of models that use Lasso/Ridge/ElasticNet regularization. We follow the procedure of GKX in this regard. 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.[10]

[10] This explains why we cannot reorder the training, validation & test sets with classical cross-validation.

Our tuning procedure is as follows: in each slice, any given model that requires tuning of some hyperparameter(s) λ is estimated multiple times on the training set for a range of possible λ values. Then the optimal value of each λ is selected based upon the performance of the model on the validation set, as measured by mean-squared error. Finally, the entire model is re-estimated on a concatenation of the training and validation sets, for the optimal value of each λ. Any models that do not require tuning of some λ hyperparameter(s) are directly estimated on the concatenation of the training and validation sets. This procedure enables us to assess each model's out-of-sample performance on the test set, per slice. We report out-of-sample performance figures per model, computed on all its test-set forecasts.

This empirical design produces true out-of-sample estimates of predictive performance, since no model is estimated in any way on any data within the test set, only assessed on this data. GKX report model performance using the following out-of-sample (OOS) predictive R^2, and we follow that study in doing likewise:

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

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.[11]

As there are no known parametric tests involving the R^2_{OOS} metric, we rely on a nonparametric bootstrap analysis to perform statistical tests on these quantities. 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. 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

[11] A similar measure is also employed by Gu et al. (2020a).

that we wish to compare, then generate predictions, from which we calculate the appropriate R^2_{OOS} quantities. In comparing two models, we are interested in the difference between the corresponding pair of R^2_{OOS} 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_{OOS} difference of interest. Note that, while 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 zero for any given significance level, much like the analysis that typically follows a classical linear regression.[12]

3.3 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. 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.

[12]Interestingly, exact post-selection statistical tests have recently been developed for Lasso-regularized models, like some of our own, notably the "fixed-λ" test of Lee et al. (2016) and the "spacing test" of Tibshirani et al. (2016). However, we found it impractical to perform the computations necessary for those two exact tests on our large dataset, for the purposes of comparing different estimates. Indeed, Lee et al. (2016, pp. 921) discuss one such practical limitation to their test. In contrast, the bootstrap, while 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. Future methodological developments might produce exact confidence intervals for general classes of (potentially nonlinear) ML models, but bootstrap-based methods have the most general applicability for now.

SIC codes

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

[Insert Table 2 about here]

A firm's industry is based on its SIC code, according to the ranges defined in Table 2. SIC codes are qualitative labels assigned by the US government according to the nature of firms' primary business activities at the time of assignment. If they are accurate[13] 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 2 because that would require us to make a further subjective decision on what level of grouping in the hierarchical SIC codes structure would be most useful.

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*[14] 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 et al. (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.

[Insert Algorithm 1 about here]

Algorithm 1 presents a basic alternating optimization procedure to perform k-means clustering.[15] Firm i data is denoted by vector x_i, the scalar z_i denotes its assignment to a cluster

[13] Since SIC codes are manually assigned and rarely changed, it is possible that they do not accurately reflect firms' activities.

[14] Throughout this paper, we use the terms "cluster" and "clustering" in relation to the technique of cluster analysis, rather than the econometric meaning of adjusting standard errors to account for dependencies.

[15] Algorithm 1 is based upon the canonical implementation of MacQueen et al. (1967). More efficient implementations are also available – in fact, we use a variant due to Hartigan and Wong (1979) – but the intuition is identical. Algorithm 1 has also been modified for other applications, see Patton and Weller (2019).

and the vector μ_k denotes the center of a cluster. The intuition behind k-means clustering (and Algorithm 1) is as follows: initialize a fixed number of clusters K at coordinates $\mu_1, \ldots, \mu_K$. Then update the cluster locations $\mu_1, \ldots, \mu_K$ to minimize the sum of within-cluster variances (the objective). Repeat these interlocking steps until the clusters no longer change: the resulting partition is the one for which within-cluster variances are smallest; i.e. within-cluster firm dissimilarities (based on observable characteristics) are minimized. In our setting, each input point x_i to the algorithm represents the scaled means of firm i's characteristics vector c_{it}. The elements of x_i do not include firm i's industry nor do they include the firm's excess return.

The number of clusters to use in each dataset 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 out-of-sample performance of those models may suffer if we pick an incorrect number of clusters; we will see in the results that this 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 z_i; rather, it aims to infer the z_i cluster memberships based on observables x_i. We therefore assume that the observables in our dataset can be used to proxy for firms' latent group memberships. In passing cluster assignments z_i to our predictive models, we are effectively combining an unsupervised learning stage with a supervised learning one.[16]

[16]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-

4 Data

4.1 Databases

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

We join this CRSP database to Compustat and I/B/E/S 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.

4.2 Characteristics

We follow GHZ in defining firm-level characteristics from the CRSP, Compustat and I/B/E/S databases. The firm characteristics in question are also known in the literature as "anomaly characteristics" because portfolios formed by sorting on such 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.[17]

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*, *tms_mkt*. This

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 towards 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.

[17] We omit the *realestate* firm-level characteristic, which GHZ included, because our sample did not include any observations prior to 1985.

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 GKX, we use more firm-level characteristics. We omit *realestate*, which GKX included. We include *chfeps*, *chnanalyst*, *disp*, *fgr5yr*, *ipo*, *nanalyst*, *sfe* & *sue*, which GKX omitted and we use the same set of market-level variables, without any interaction terms.[18]

4.3 Pre-processing

We follow GKX in a number of key data pre-processing steps. Firm-level characteristics are rescaled to the range $[-1, +1]$ cross-sectionally; i.e. 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.

While 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,[19] shifting the start date forward even further. We begin at a year where all characteristics are widely available.

5 Empirical Results

5.1 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. Our method of selecting

[18] Refer to the Internet Appendix for details of how the firm-level and market-level variables have been constructed.

[19] *grcapx* requires 2 years of prior data. *beta* and *idiovol* each require 3 years of prior returns.

the optimal number of clusters consistently returns 2 to 4 clusters, no matter whether we apply it to the top 1,000 (by market capitalization), top 2,000, or all available firms in our sample.

[Insert Figure 1 about here]

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 from is a welcome feature when interpreting these clusters in terms of characteristics.

Interpreting the clusters

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

[Insert Figure 2 about here]

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 a 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 three 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.

5.2 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 out-of-sample predictability. To show this, we use the industry partition of the cross-section to estimate pooled models, as defined in Eqn. (5), as well as by-industry models, as defined in Eqn. (8), in order to compare their out-of-sample predictive performance. We employ a variety of regularization schemes to deal with the challenging nature of this high-dimensional dataset, and also report results from unregularized OLS models.

[Insert Table 3 about here]

Table 3 reports the attained out-of-sample predictability results by industry. Other than unregularized OLS models, the out-of-sample 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.

[Insert Table 4 about here]

Table 4 shows the *incremental* out-of-sample predictability of by-industry over pooled models; each cell contains the R^2_{OOS} (%) of a by-group model minus the R^2_{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 & mining), the incremental out-of-sample 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 different to zero are also positive. This statistical evidence indicates that allowing heterogeneous predictive relationships across industries can positively impact out-of-sample predictability at the firm level.

We saw in Section 5.1 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 out-of-sample predictability at the firm level.

[Insert Table 5 about here]

Table 5 reports the attained out-of-sample predictability results by cluster. Once again, and besides the unregularized OLS models, the out-of-sample 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.

[Insert Table 6 about here]

Table 6 shows the *incremental* out-of-sample 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 & 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 to zero. Therefore, for Lasso and ElasticNet-regularized models, this statistical evidence indicates that allowing heterogeneous predictive relationships across clusters can also positively impact out-of-sample predictability at the firm level.

5.3 Predictive performance with two-stage procedure

Following the prior literature on firm-level return predictability, we now assess out-of-sample predictive performance on the full cross-section of firms, rather than by industry or by cluster. In doing so, we follow the procedure laid out in Eqns. (9) through (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.

Table 7 shows the aggregate out-of-sample 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 shows 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.

[Insert Table 7 about here]

Many of the models exhibit a positive out-of-sample R^2_{OOS}, in spite of the difficulty of predicting firm-level returns. The best-performing model in each case is a Ridge-regularized two-stage 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 towards one another (Hastie et al., 2009). It is worth noting that not only does OLS perform poorly, but attempting to incorporate heterogeneity in predictive relationships even worsens the predictive performance. This highlights the need for regularization in our high-dimensional setting.

Table 8 shows the aggregate out-of-sample 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 out-of-sample R^2_{OOS} values.

[Insert Table 8 about here]

The headline result from Table 8 Panel (c) is that the Two-stage Lasso model performs best overall when estimating on and predicting for all available firms. It also exhibits the best individual performance for each of the three 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-based models now perform the best, rather than Ridge-based ones. This facilitates an interpretation of which characteristics matter for predictability, as we shall see in Section 6.

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 out-of-sample predictability.

Comparison with existing results

The closest framework is that of GKX; indeed, we adopt a similar slicing strategy and the same measure of out-of-sample performance (R^2_{OOS}) in order to facilitate comparisons between our results. Table 1 in GKX indicates that the best neural network methods detailed in that paper have an $R^2_{OOS} = 0.40\%$ when evaluated on their full sample.

To draw a contrast with some of our own headline figures, some variants of our two-stage predictive strategy in Table 8 Panel (a) exhibit an out-of-sample performance level of $R^2_{OOS} > 1.9\%$ when we train on a subset of the largest firms only and exploit cluster heterogeneity. Procedure-wise, our most comparable results to GKX are those where we do not restrict ourselves to clustered firms: in Table 7 Panel (c), our best (linear) model exploits heterogeneity to reach an out-of-sample performance level of $R^2_{OOS} = 0.76\%$. When we use the clustering procedure, in Table 8 Panel (c), our best (linear) model achieves an $R^2_{OOS} = 1.05\%$. Any further comparisons are constrained by our different samples. 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.

It is also possible to 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." Furthermore, our R^2_{OOS} metric is an out-of-sample statistic, and Rapach and Zhou (2013) note that "out-of-sample R^2 statistics will frequently be even lower", implying that out-of-sample values around 1% are even more economically notable than in-sample values.

6 Variable importance and heterogeneity

Prior studies (Gu et al., 2020b; Freyberger et al., 2020; DeMiguel et al., 2020) took various approaches to answer Cochrane (2011)'s 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 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 5.1, 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[20] in forming those clusters of firms: firm *age* (and the related measure of whether they recently IPO'd), *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-based models to discard irrelevant predictive characteristics for each cluster.

[Insert Table 9 about here]

Table 9 displays the frequency with which firm-level predictive characteristics are selected for each cluster of firms using the by-cluster Lasso procedure. Much as we used similar models to find evidence of heterogeneity in predictability earlier, we now exploit the by-cluster Lasso model to see that the selected predictive characteristics do indeed vary by cluster. This interpretability does not come at the expense of performance because the Lasso-based models have the best out-of-sample predictive performance.

[Insert Table 10 about here]

The use of a Lasso penalty selects a subset of coefficients by setting all non-selected coefficients' values to zero, as we see in Table 9. We go further and test whether the coefficients are significantly different to zero by computing bootstrap confidence intervals for each coefficient. Table 10 displays the results of such a bootstrap analysis: it shows the number of coefficients (displayed as a frequency) that are non-zero at a 99% level of significance. By comparing Tables 9 and 10, 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*, *dp_mkt*)

[20] We conduct an alternative analysis of characteristic importance for cluster formation in the Internet Appendix, which gives similar results.

are never significantly non-zero for Cluster 3 even though they were selected by the Lasso for a third of our slices.

[Insert Table 11 about here]

Table 11 displays the frequency of selection of characteristics by the Two-Stage Lasso 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, price reversal), market beta, book-to-market and earnings-to-price.

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). 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 (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*, *cashpr*) and the market-level D/P ratio (*dp_mkt*) tend to vary across clusters of firms, while a few other variables (*baspread*, *roic*, *mve* & *sue*) are only selected at the level that is common to all firms in the cross-section. The good out-of-sample performance of Lasso-based models that use only these selected coefficients confirms their importance. This parsimony echoes Kelly et al. (2019), who found that only a small subset of stock characteristics were 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.

Another contribution is providing statistical evidence for the importance of our selected predictive variables using the bootstrap, in contrast to the non-statistical measures used in the prior literature.

7 Conclusion

We build on GKX who state: "Machine learning methods on their own do not identify deep fundamental associations among asset prices and conditioning variables. When the objective is to understand economic mechanisms, machine learning still may be useful. It requires the economist to add structure – to build a hypothesized mechanism into the estimation problem – and decide how to introduce a machine learning algorithm subject to this structure".

We make progress on several fronts by developing 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 upon observable characteristics. Our procedure incorporates an economically-motivated notion of heterogeneity in the cross-section.

Allowing for heterogeneity in predictive relationships favorably impacts our empirical ability to estimate next-period returns. To select which characteristics matter, we use interpretable sparse linear models to discard irrelevant predictive variables. We test which characteristics were most frequently selected by our sparse models, and thus most important in determining next-period returns.

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 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 (Primiceri et al., 2018) 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. Indeed, sparsity has proven useful for model selection (Feng et al., 2020), empirical causal inference (Belloni et al., 2012, 2014; Chernozhukov et al., 2017) and theoretical modeling of economic behavior (Gabaix, 2014; Hanna et al., 2014; Croce et al., 2015; Gabaix, 2020; Gueciouer, 2020), where it can be related to broader forms of inattention (Sims, 2003; Reis, 2006; Chetty et al., 2009; Gabaix, 2019).

References

Ando, T., Bai, J., 2017. 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, 1182–1198.

Argyriou, A., Evgeniou, T., Pontil, M., 2006. Multi-task feature learning. In: *Advances in Neural Information Processing Systems*, pp. 41–48.

Asness, C. S., Porter, R. B., Stevens, R. L., 2000. Predicting stock returns using industry-relative firm characteristics. Available at SSRN 213872.

Barrot, J.-N., Sauvagnat, J., 2016. Input specificity and the propagation of idiosyncratic shocks in production networks. Quarterly Journal of Economics 131, 1543–1592.

Belloni, A., Chen, D., Chernozhukov, V., Hansen, C., 2012. Sparse models and methods for optimal instruments with an application to eminent domain. Econometrica 80, 2369–2429.

Belloni, A., Chernozhukov, V., Hansen, C., 2014. Inference on treatment effects after selection among high-dimensional controls. Review of Economic Studies 81, 608–650.

Bonhomme, S., Manresa, E., 2015. Grouped patterns of heterogeneity in panel data. Econometrica 83, 1147–1184.

Campbell, J. Y., Thompson, S. B., 2008. Predicting excess stock returns out of sample: Can anything beat the historical average? Review of Financial Studies 21, 1509–1531.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., 2017. Double/debiased/Neyman machine learning of treatment effects. American Economic Review Papers and Proceedings 107, 261–65.

Chetty, R., Looney, A., Kroft, K., 2009. Salience and taxation: Theory and evidence. American Economic Review 99, 1145–77.

Cochrane, J. H., 2011. Presidential address: Discount rates. Journal of Finance 66, 1047–1108.

Cohen, L., Frazzini, A., 2008. Economic links and predictable returns. Journal of Finance 63, 1977–2011.

Croce, M. M., Lettau, M., Ludvigson, S. C., 2015. Investor information, long-run risk, and the term structure of equity. Review of Financial Studies 28, 706–742.

Daniel, K., Mota, L., Rottke, S., Santos, T., 2020. The cross-section of risk and returns. Review of Financial Studies 33, 1927–1979.

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

Diebold, F. X., Shin, M., 2019. Machine learning for regularized survey forecast combination: Partially-egalitarian lasso and its derivatives. International Journal of Forecasting 35, 1679–1691.

Ding, C., He, X., 2004. K-means clustering via principal component analysis. In: *Proceedings of the twenty-first International Conference on Machine Learning*, p. 29.

Dorn, D., Huberman, G., 2010. Preferred risk habitat of individual investors. Journal of Financial Economics 97, 155–173.

Evgeniou, T., Pontil, M., 2004. Regularized multi–task learning. In: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 109–117.

Feng, G., Giglio, S., Xiu, D., 2020. Taming the factor zoo: A test of new factors. Journal of Finance 75, 1327–1370.

Fisher, J., Puelz, D., Carvalho, C. M., 2020. Monotonic effects of characteristics on returns. Annals of Applied Statistics (forthcoming) .

Freyberger, J., Neuhierl, A., Weber, M., 2020. Dissecting characteristics nonparametrically. Review of Financial Studies 33, 2326–2377.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., Walther, A., 2020. Predictably unequal? The effects of machine learning on credit markets. Journal of Finance (forthcoming) .

Gabaix, X., 2014. A sparsity-based model of bounded rationality. Quarterly Journal of Economics 129, 1661–1710.

Gabaix, X., 2019. Behavioral inattention. In: *Handbook of Behavioral Economics: Applications and Foundations*, Elsevier, vol. 2, pp. 261–343.

Gabaix, X., 2020. A behavioral New Keynesian model. American Economic Review 110, 2271–2327.

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

Grishchenko, O. V., Rossi, M., 2012. The role of heterogeneity in asset pricing: The effect of a clustering approach. Journal of Business & Economic Statistics 30, 297–311.

Gu, S., Kelly, B., Xiu, D., 2020a. Autoencoder asset pricing models. Journal of Econometrics (in press).

Gu, S., Kelly, B., Xiu, D., 2020b. Empirical asset pricing via machine learning. Review of Financial Studies 33, 2223–2273.

Gueciouer, A., 2020. How do investors learn as data becomes bigger? Evidence from a FinTech platform. Available at SSRN 3708476.

Han, Y., He, A., Rapach, D., Zhou, G., 2020. Firm characteristics and expected stock returns. Available at SSRN 3185335.

Hanna, R., Mullainathan, S., Schwartzstein, J., 2014. Learning through noticing: Theory and evidence from a field experiment. Quarterly Journal of Economics 129, 1311–1353.

Hartigan, J. A., Wong, M. A., 1979. Algorithm AS 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics) 28, 100–108.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media, second ed.

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

Hou, K., 2007. Industry information diffusion and the lead-lag effect in stock returns. Review of Financial Studies 20, 1113–1138.

Hou, K., Robinson, D. T., 2006. Industry concentration and average stock returns. Journal of Finance 61, 1927–1956.

Jalali, A., Sanghavi, S., Ruan, C., Ravikumar, P. K., 2010. A dirty model for multi-task learning. In: *Advances in Neural Information Processing Systems*, pp. 964–972.

Jiang, G., Lee, C. M., Zhang, Y., 2005. Information uncertainty and expected returns. Review of Accounting Studies 10, 185–221.

Karolyi, G. A., Van Nieuwerburgh, S., 2020. New methods for the cross-section of returns. Review of Financial Studies 33, 1879–1890.

Kelly, B. T., Pruitt, S., Su, Y., 2019. Characteristics are covariances: A unified model of risk and return. Journal of Financial Economics 134, 501–524.

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

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

Lee, J. D., Sun, D. L., Sun, Y., Taylor, J. E., 2016. Exact post-selection inference, with application to the lasso. Annals of Statistics 44, 907–927.

Lewellen, J., 1999. The time-series relations among expected return, risk, and book-to-market. Journal of Financial Economics 54, 5–43.

Lewellen, J., 2015. The cross-section of expected stock returns. Critical Finance Review 4, 1–44.

MacQueen, J., et al., 1967. Some methods for classification and analysis of multivariate observations. In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Oakland, CA, USA, vol. 1, pp. 281–297.

Menzly, L., Ozbas, O., 2010. Market segmentation and cross-predictability of returns. Journal of Finance 65, 1555–1580.

Menzly, L., Santos, T., Veronesi, P., 2004. Understanding predictability. Journal of Political Economy 112, 1–47.

Moskowitz, T. J., Grinblatt, M., 1999. Do industries explain momentum? Journal of Finance 54, 1249–1290.

Patton, A. J., Weller, B., 2019. Risk price variation: The missing half of empirical asset pricing. Tech. rep., Economic Research Initiatives at Duke (ERID) Working Paper No. 274.

Primiceri, G. E., Lenza, M., Giannone, D., et al., 2018. Economic predictions with big data: The illusion of sparsity. Tech. rep., Federal Reserve Bank of New York Staff Report No. 847.

Rapach, D., Zhou, G., 2013. Forecasting stock returns. In: *Handbook of Economic Forecasting*, vol. 2, pp. 328–383.

Rapach, D. E., Strauss, J. K., Tu, J., Zhou, G., 2019. Industry return predictability: A machine learning approach. Journal of Financial Data Science 1, 9–28.

Rapach, D. E., Strauss, J. K., Zhou, G., 2010. Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. Review of Financial Studies 23, 821–862.

Rapach, D. E., Zhou, G., 2020. Time-series and Cross-sectional Stock Return Forecasting: New Machine Learning Methods, Wiley, chap. 1, pp. 1–33.

Reis, R., 2006. Inattentive consumers. Journal of Monetary Economics 53, 1761–1800.

Ross, S. A., 2005. Neoclassical Finance. Princeton University Press.

Rousseeuw, P. J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20, 53–65.

Santos, T., Veronesi, P., 2006. Labor income and predictable stock returns. Review of Financial Studies 19, 1–44.

Sims, C. A., 2003. Implications of rational inattention. Journal of Monetary Economics 50, 665–690.

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

Tibshirani, R. J., Taylor, J., Lockhart, R., Tibshirani, R., 2016. Exact post-selection inference for sequential regression procedures. Journal of the American Statistical Association 111, 600–620.

Tikhonov, A. N., Arsenin, V. Y., 1977. Solutions of ill-posed problems. Winston.

Tropp, J. A., 2006. Just relax: Convex programming methods for identifying sparse signals in noise. IEEE transactions on Information Theory 52, 1030–1051.

Wainwright, M. J., 2009. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). IEEE transactions on Information Theory 55, 2183–2202.

Welch, I., Goyal, A., 2007. A comprehensive look at the empirical performance of equity premium prediction. Review of Financial Studies 21, 1455–1508.

Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68, 49–67.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Methodological) 67, 301–320.

Appendix

Table 1: Time periods covered by each slice.

Slice	Training	Validation	Test
1	1984-2003	2004-2009	2010
2	1984-2004	2005-2010	2011
3	1984-2005	2006-2011	2012
4	1984-2006	2007-2012	2013
5	1984-2007	2008-2013	2014
6	1984-2008	2009-2014	2015

Table 2: 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 & utilities	4011–4991
wholesale	5000–5190

Algorithm 1 Pseudocode for the K-means clustering algorithm

1: Initialise K clusters
2: **repeat**
3: (Re)assign each observation to the closest (in squared Euclidean distance) cluster mean: $z_i = \arg\min_k ||x_i - \mu_k||^2$
4: Update the means of the currently assigned clusters: $\mu_k = \frac{1}{N_k} \sum_{i:z_i=k} x_i$
5: **until** convergence

Table 3: Industry-level out-of-sample predictability, measured by R^2_{OOS} (%), when partitioning firms by industry.

Model	agriculture	construction	finance	manufacturing	mining	noclassif	retail	services	transport & utilities	wholesale
By-industry Lasso	0.82	1.40	0.65	0.73	0.34	1.59	0.53	0.70	0.61	0.86
By-industry ElasticNet	0.70	1.60	0.76	0.75	0.31	1.94	0.46	0.80	0.69	0.97
By-industry Ridge	0.83	1.61	0.81	0.78	0.32	1.97	0.48	0.85	0.74	1.02
By-industry OLS	-9.26	-6.58	-5.99	-4.79	-6.97	0.91	-5.58	-5.62	-3.24	-5.24
Pooled Lasso	1.02	0.98	0.37	0.36	0.47	0.34	0.25	0.32	0.28	0.36
Pooled ElasticNet	1.05	1.04	0.32	0.26	0.44	0.22	0.10	0.25	0.17	0.31
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

Note: The models are each estimated based on all available firms, once per slice.

Table 4: Incremental industry-level out-of-sample predictability, measured by R^2_{OOS} (%), when partitioning firms by industry. The increments are calculated as the difference between by-industry and pooled R^2_{OOS}, for each regularization scheme. Asterisks indicate that the increments are significantly different to zero, based on bootstrap confidence intervals.

Regularization	agriculture	construction	finance	manufacturing	mining	noclassif	retail	services	transport & 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

Note: Significance tests are based on 1,000 bootstrap samples. Asterisks denote the significance level (***=99%, **=95%, *=90%).

Table 5: Cluster-level out-of-sample predictability, measured by R^2_{OOS} (%), when partitioning firms by cluster.

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

Note: 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.

Table 6: Incremental cluster-level out-of-sample predictability, measured by R^2_{OOS} (%), when partitioning firms by cluster. The increments are calculated as the difference between by-cluster and pooled R^2_{OOS}, for each regularization scheme. Asterisks indicate that the increments are significantly different to zero, based on bootstrap confidence intervals.

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 ***

Note: Significance tests are based on 1,000 bootstrap samples. Asterisks denote the significance level (***=99%, **=95%, *=90%).

Table 7: Aggregate out-of-sample predictability, measured by R^2_{OOS} (%), when partitioning firms by industry.

Panel (a)

Model	Top 1,000
Two-stage Ridge	1.65
By-industry Ridge	1.60
Pooled ElasticNet	1.57
Pooled Ridge	1.57
By-industry ElasticNet	1.54
Pooled Lasso	1.52
By-industry Lasso	1.49
Two-stage Lasso	1.49
Pooled OLS	-8.70
By-industry OLS	-14.78
Two-stage OLS	-14.78

Panel (b)

Model	Top 2,000
Two-stage Ridge	1.38
By-industry Ridge	1.34
Pooled ElasticNet	1.33
Pooled Ridge	1.33
By-industry ElasticNet	1.31
By-industry Lasso	1.29
Pooled Lasso	1.29
Two-stage Lasso	1.29
Pooled OLS	-7.36
By-industry OLS	-12.05
Two-stage OLS	-12.05

Panel (c)

Model	All Firms
Two-stage Ridge	0.76
Pooled ElasticNet	0.73
Pooled Ridge	0.73
By-industry Ridge	0.72
Pooled Lasso	0.71
By-industry ElasticNet	0.69
By-industry Lasso	0.65
Two-stage Lasso	0.65
Pooled OLS	-3.77
By-industry OLS	-5.44
Two-stage OLS	-5.44

Note: Each panel represents results from estimating the models based on a particular subset of results: (a) on the largest 1,000 firms by market capitalization, (b) on the largest 2,000 firms, and (c) on the full sample.

Table 8: Aggregate out-of-sample predictability, measured by R^2_{OOS} (%), when partitioning firms by cluster.

Panel (a)

Model	Top 1,000
Two-stage Ridge	1.91
Pooled Ridge	1.91
Pooled Lasso	1.88
By-cluster Ridge	1.86
Pooled ElasticNet	1.85
By-cluster ElasticNet	1.83
Two-stage Lasso	1.78
By-cluster Lasso	1.77
Pooled OLS	-8.86
By-cluster OLS	-30.86
Two-stage OLS	-30.86

Panel (b)

Model	Top 2,000
By-cluster Lasso	1.61
Two-stage Lasso	1.61
Pooled Lasso	1.60
By-cluster ElasticNet	1.59
Pooled Ridge	1.58
Two-stage Ridge	1.58
By-cluster Ridge	1.55
Pooled ElasticNet	1.53
Pooled OLS	-8.23
By-cluster OLS	-20.92
Two-stage OLS	-20.92

Panel (c)

Model	All Firms
Two-stage Lasso	1.05
By-cluster Lasso	1.03
By-cluster ElasticNet	1.03
Pooled Lasso	0.97
Two-stage Ridge	0.96
By-cluster Ridge	0.95
Pooled Ridge	0.95
Pooled ElasticNet	0.94
Pooled OLS	-4.81
By-cluster OLS	-61.38
Two-stage OLS	-61.38

Note: Each panel represents results from estimating the models based on a particular subset of results: (a) on the largest 1,000 firms by market capitalization, (b) on the largest 2,000 firms, and (c) on the full sample.

Table 9: Frequency of selection (% of slices) of characteristics by cluster, when estimating the By-cluster Lasso model.

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

Note: The model was estimated based on all available firms, once per slice.

Table 10: Frequency (% of slices) for which each coefficient is significant (at the 99% level) by cluster, when estimating the By-cluster Lasso model.

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

Note: Significance tests are based on 1,000 bootstrap samples.

Table 11: Frequency of selection (% of slices) of characteristics by cluster, when estimating the Two-stage Lasso model.

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

Note: The model was estimated based on all available firms, once per slice.

Figure 1: Visualisations 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.

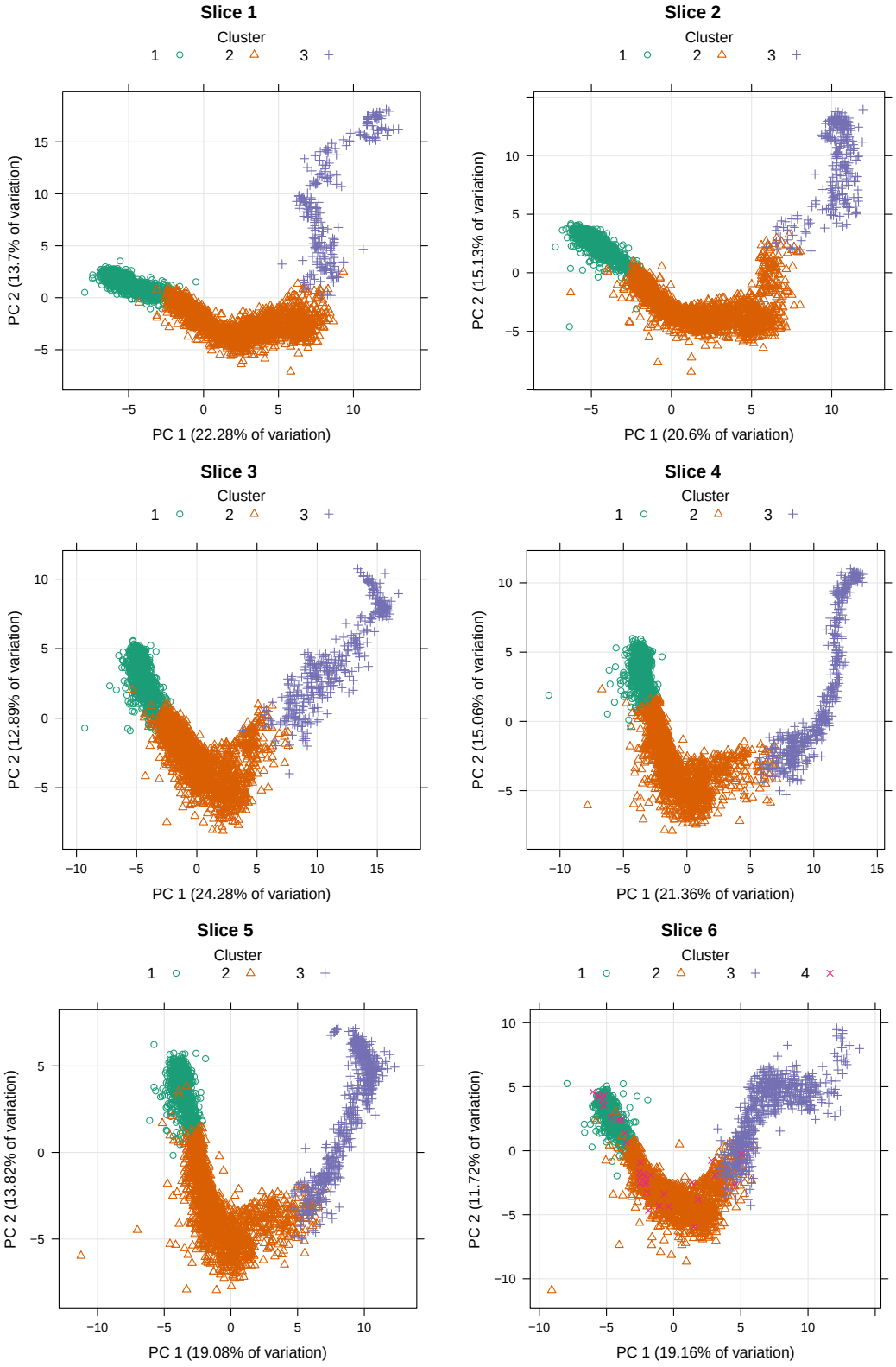
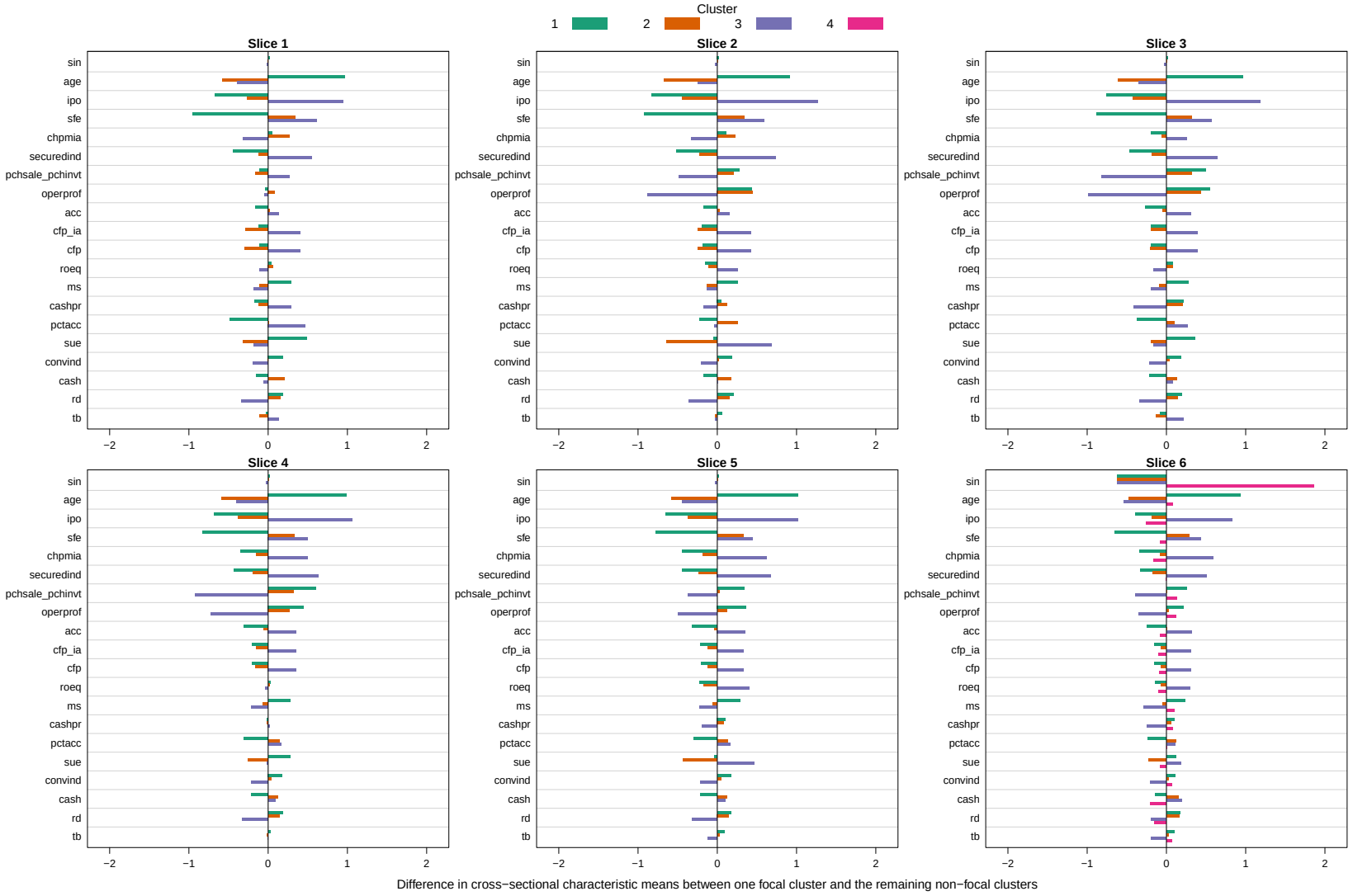


Figure 2: Interpreting clusters by comparing their characteristic means. Each bar in this figure depicts the difference between a cluster's (cross-sectional) characteristic mean and the characteristic mean calculated for all other clusters. The 20 firm-level characteristics with the highest such absolute 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 visualises a single slice's clusters.



Internet Appendix

A Characteristics

Tables 12 and 13 detail the market-level and firm-level (respectively) predictive variables used in this study. Refer to Welch and Goyal (2007) and Green et al. (2017) (respectively) for further details on each.

Table 12: Full list of market-level variables used in our study. All are at a monthly frequency. Refer to Welch and Goyal (2007) for further details.

Code	Name	Type	Description
bm_mkt	Book-to-Market	ratio	Ratio of book value to market value for the Dow Jones Industrial Average
dfy_mkt	Default Spread Yield	rate	Difference between BAA and AAA-rated corporate bond yields
dp_mkt	Dividend-Price Ratio	ratio	Difference between log dividends (12-month moving sum) and log prices for the S&P500
ep_mkt	Earnings-Price Ratio	ratio	Difference between log earnings (12-month moving sum) and log prices for the S&P500
ntis_mkt	Net Equity Expansion	ratio	12-month moving sums of net issues/end-of-year market cap for NYSE stocks
svar_mkt	Stock Variance	rate	Sum of squared daily returns for the S&P 500
tbl_mkt	Treasury bill rate	rate	3-month US Treasury bill rates
tms_mkt	Term Spread	rate	Difference between the long-term yield on government bonds and the Treasury bill

Table 13: Full list of firm-level characteristics used in our study. Refer to Green et al. (2017) for further details.

Code	Name	Frequency	Type	Description
absacc	Absolute Value of Accruals, scaled by AT	Yearly	ratio	Absolute Value of Working Capital Accruals, scaled by AT
acc	Accruals, scaled by AT	Yearly	ratio	Working Capital Accruals, scaled by AT
aeavol	Abnormal volume around earnings announcement	Quarterly	ratio	Average volume 3 days around earnings announcement relative to 10-30 day window before announcement, scaled by monthly volume
age	Years of Coverage	Yearly	number of years	Years since First Compustat Coverage
agr	% change in assets	Yearly	percentage	Annual percentage change in assets (at)
baspread	Bid-ask spread	Monthly	ratio	Monthly average of (daily bid-ask spread divided by average of daily spread)
beta	Market Beta	Monthly	coefficient	Market beta based on 36 months of weekly returns
betasq	Market Beta Squared	Monthly	coefficient	Market beta squared based on 36 months of weekly returns
bm	book-to-market	Yearly	ratio	Book value of equity (ceq) divided by end of fiscal-year market capitalization
bm_ia	SIC2-adj. book-to-market	Yearly	ratio (adjusted)	Industry-Adjusted Book value of equity (ceq) divided by end of fiscal-year market capitalization
cash	Cash Holdings	Quarterly	ratio	Cash and cash equivalents divided by average total assets
cashdebt	Cash flow to debt	Yearly	ratio	Earnings before depreciation and extra items over avg. liabilities
cashpr	Cash Productivity	Yearly	ratio	Fiscal year-end market cap plus long term debt minus total assets divided by cash and equiv assets
cfp	Cash Flow to Price Ratio	Yearly	ratio	Operating cash flows/fiscal-year-end market capitalization
cfp_ia	SIC2-adj. Cash Flow to Price Ratio	Yearly	ratio	SIC2-adj. operating cash flows/fiscal-year-end market capitalization
chatoia	SIC2-adj. change in asset turnover	Yearly	change in ratio	The 2-digit SIC fiscal-year mean adjusted change in sales divided by average total assets
chcsho	% change in shares outstanding	Yearly	percentage	Annual percentage change in shares outstanding (csho)

Code	Name	Frequency	Type	Description
chempia	SIC2-adj. % change in employees	Yearly	percentage	Industry adjusted annual percent change in employees (hire)
chfeps	Change in Forecasted (mean) EPS	Monthly	change	1-month change in forecasted mean EPS
chinv	Change in Inventory	Yearly	ratio	Change in inventory scaled by average total assets
chmom	Change in 6-month-momentum	Monthly	percent	Cumulative returns from months t-6 to t-1 minus months t-12 to t-7
chnanalyst	Change in Number of Analysts	Monthly	change	3-month change in number of analysts
chpmia	SIC2-adj. change in profit margin	Yearly	(change in) ratio	Industry-adjusted annual change in profit margin
chtx	Change in 4-quarter tax expense	Quarterly	ratio	Change in total taxes from quarter t-4, scaled by total assets (t-4)
cinvest	Corporate Investment	Quarterly	ratio	Change in net PP&E over sales net of mean of this over prior 3 quarters
convind	Convertible Debt Indicator	Yearly	dummy	Indicator for whether company has convertible debt obligations
currat	Current Ratio	Yearly	ratio	Current assets/current liabilities
depr	Depreciation/PP&E	Yearly	ratio	Depreciation/PP&E
disp	Dispersion in Forecasts	Monthly	Ratio	Standard deviation of analysts' forecasts over mean forecast
divi	Dividend Initiation	Yearly	dummy	1 if company pays dividends but did not in prior year
divo	Dividend Omission	Yearly	dummy	1 if company does not pay dividend but did in prior year
dolvol	Dollar trading volume in month t-2	Monthly	dollar value	2-month lag of volume times price
dy	Dividends to Price	Yearly	ratio	Total dividends (dvt) divided by market capitalization at fiscal year-end
ear	Sum daily returns 3 days around earnings announcement	Quarterly	sum	Sum of daily returns in three days around earnings announcement
egr	% Change in common shareholder equity	Yearly	percentage	Annual percent change in book value of equity (ceq)

Code	Name	Frequency	Type	Description
ep	Earnings to Price	Yearly	ratio	Annual income before extraordinary items (ib) divided by end of fiscal year market cap
fgr5yr	Forecasted 5-year growth	Monthly	percentage	Most recently available analyst forecasted 5-year growth
gma	Gross profitability	Yearly	ratio	Revenues (revt) minus cost of goods sold (cogs) divided by lagged total assets (at)
grcapx	2-year growth of Cap. Expenditures	Yearly	percentage	Percent change in capital expenditures from year t-2 to year t
grltnoa	Growth in Long-Term Net Operating Assets	Yearly	ratio	Growth in Long-Term Net Operating Assets
herf	Herfindahl index	Yearly	percentage	2-digit SIC - fiscal-year sales concentration (sum of squared percent of sales in industry for each company)
hire	% change in employees	Yearly	percentage	Industry-adjusted annual percent change in employees (hire)
idiovol	Idiosyncratic return volatility	Monthly	regression estimate	Standard deviation of residuals from regressions of weekly returns on equal weighted market returns for 3 years
ill	Illiquidity	Monthly	ratio	Monthly average of daily (absolute return / dollar volume)
indmom	12-month-momentum industry average	Monthly	percent	12-month-momentum by industry
invest	Capital Expenditures & Inventory	Yearly	ratio	Annual change in PPEGT + annual change in inventories scaled by lagged total assets
ipo	IPO indicator	Monthly	dummy	Dummy if it's the first year PERMNO is available on CRSP monthly stock file
lev	Leverage	Yearly	ratio	Total liabilities (lt) divided by fiscal year-end market capitalization
lgr	% Change in long-term debt	Yearly	percentage	Annual percent change in total liabilities (lt)
maxret	Maximum Daily Return	Monthly	Return	Maximum daily return from month t-1
mom12m	12-month-momentum	Monthly	percent	12-month-momentum
mom1m	1-month-momentum	Monthly	percent	1-month-momentum
mom36m	36-month-momentum	Monthly	percent	Cumulative returns from months t -36 to t - 13

Code	Name	Frequency	Type	Description
mom6m	6-month-momentum	Monthly	percent	6-month-momentum
ms	Financial Statement (Moharan) Score	Yearly	score out of 8	Sum of 8 indicator variables (quarterly and annual)
mve	Market capitalization	Monthly	dollar value	Natural log of market capitalization at end of month t-1
mve_ia	SIC2-adj. firm size	Yearly	dollar value	2-digit SIC industry-adjusted fiscal year-end market capitalization
nanalyst	Analyst Count	Monthly	integer	Most recently available number of analysts following stock
nincr	Number of earnings increases in most recent 8 quarters	Quarterly	numeric	Number of consecutive quarters (up to eight quarters) with an increase in earnings (ibq)
operprof	Operating Profitability	Yearly	ratio	Revenue - cost goods sold - SG&A expense - interest expense over lagged common equity
orgcap	Organizational Capital	Yearly	ratio	Capitalized SG&A expenses (annual)
pchcapx_ia	SIC2-adj. % change in capital expenditures	Yearly	percentage	The 2-digit SIC fiscal-year mean adjusted change in capital expenditures
pchcurrat	% Change in Current Ratio	Yearly	percentage	% change in current assets/current liabilities
pchdepr	% Change in Depreciation	Yearly	percentage	% change in depreciation
pchgm_pchsale	% change gross margin - % change sales	Yearly	percentage	% change in gross margin minus percent change in sales
pchquick	% change in Quick Ratio	Yearly	percentage	% change in (current assets - inventory) / current liabilities
pchsale_pchinvt	% Ch.Sales - % Ch.Inventory	Yearly	percentage	% change in sales - % Change in inventory
pchsale_pchrect	%change sales - %change receivables	Yearly	percentage	Annual percent change in sales minus annual percent change in receivables
pchsale_pchxsga	% Ch.Sales - % Ch.SG&A	Yearly	percentage	% change in sales - % Change in SG&A
pchsaleinv	% Change in Sales-to-Inventory	Yearly	percentage	% change in (sales/inventory)

Code	Name	Frequency	Type	Description
pctacc	Percent Accruals, scaled by IB	Yearly	ratio	Working capital accruals, scaled by IB
pricedelay	Price Delay	Monthly	ratio	Proportion of variation in weekly returns for 36 months ending in month t explained by 4 lags of weekly market returns (Rsquared)
ps	Financial-statements score	Yearly	score	Financial-statements score: sum of 9 indicator variables to form fundamental health score
quick	Quick Ratio	Yearly	ratio	(current assets - inventory) / current liabilities
rd	R&D increase	Yearly	dummy	Positive R&D growth relative to total assets > 5%
rd_mve	R&D to market cap	Yearly	ratio	R&D expense divided by end-of-fiscal-year market capitalization
rd_sale	R&D to sales	Yearly	ratio	R&D expense divided by sale
retvol	return volatility in month t-1	Monthly		Volatility of daily returns in month t-1
roaq	Return on Asset	Quarterly	ratio	Income (before extr. items) over 1-quarter lagged total assets
roavol	16-Quarter Earnings Volatility	Quarterly	s.d.	Standard deviation for 16 quarters of income (before extr. items) over lag total assets
roeq	Return on Equity	Quarterly	ratio	Earnings before extraordinary items divided by lagged common shareholders equity
roic	Return on Invested Capital	Yearly	ratio	Return on Invested Capital
rsup	Revenue Surprise	Quarterly	ratio	4-quarter change in sales divided by fiscal-year-end market cap
salecash	Sales-to-cash	Yearly	ratio	Annual sales divided by cash and cash equivalents
saleinv	Sales-to-inventory	Yearly	ratio	Annual sales divided by total inventory
salerec	Sales-to-receivables	Yearly	ratio	Annual sales divided by accounts receivable
secured	Secured Debt	Yearly	ratio	Total liability over secured debt
securedind	Secured Debt Indicator	Yearly	dummy	Indicator for whether company has secured debt obligations
sfe	Analyst mean annual earnings forecast (scaled)	Quarterly	ratio	Analyst mean annual earnings forecast scaled by absolute price per share at fiscal quarter end

Code	Name	Frequency	Type	Description
sgr	% growth of sales	Yearly	percentage	Annual percent change in sales
sin	Sin Stocks	Yearly	dummy	Company's primary industry is smoke or tobacco, beer or alcohol, or gaming
sp	Sales to Price	Yearly	ratio	Annual revenue (sale) divided by fiscal year-end market capitalization
std_dolvol	Volatility of dollar trading volume	Monthly	s.d.	Monthly standard deviation of daily dollar trading volume
std_turn	Volatility of share turnover	Monthly	s.d.	Monthly standard deviation of daily share turnover
stdacc	Accrual Volatility	Quarterly	s.d.	Standard deviation for 16 quarters of accruals
stdcf	Cashflow Volatility	Quarterly	s.d.	Standard deviation for 16 quarters of income (before extr. items) over lag total assets
sue	Unexpected Quarterly Earnings	Quarterly	ratio	Unexpected quarterly earnings (actual-medest I/B/E/S earnings) divided by fiscal-quarter end market cap
tang	Debt capacity/firm tangibility	Yearly	ratio	Debt capacity/firm tangibility
tb	SIC2-adj. Tax Income to Book Income	Yearly	ratio	(Tax Expense/Federal taxrate)/(income before extraordinary items)
turn	Share Turnover	Monthly	ratio	Avg 3-month trading volume/sharesout
zerotrade	Zero Trading Days	Monthly	ratio	Turnover weighted number of zero trading days for most recent month

B Details on cluster formation

This appendix provides some further details on clusters.

B.1 Firm distributions across clusters

As a first indication of how the inferred clusters of firms vary from one slice to another, Table 14 tabulates the proportions of firms that belong to each cluster. Only the 4th cluster (that appears in the final slice) appears concentrated on a small subset of firms.

Table 14: Firms (%) per cluster, for each slice.

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**
Slice 1	38.79	53.56	7.64	
Slice 2	36.58	55.36	8.06	
Slice 3	35.09	53.21	11.70	
Slice 4	33.22	51.60	15.18	
Slice 5	31.81	51.00	17.18	
Slice 6	31.08	47.88	19.94	1.09

B.2 Cluster vs. industry membership

Recall that a firm may belong to exactly one cluster and exactly one industry, for any given slice of data. Tables 15 and 16 characterize firms' joint membership of industries and clusters.

It is clear from the joint memberships that clusters do not span industries. SIC codes (from which we specify industry membership) and clusters (which are inferred from firm characteristics) are both characterizations of firms' economic activities. It is interesting to note that they appear to define different partitions of the cross-section of firms.

B.3 Importance of characteristics in cluster formation

We have provided an interpretation of the clusters in terms of their compositions in Section 5.1. We now provide an alternative interpretation of characteristic importance during the k-means clustering process, with similar conclusions.

The k-means algorithm considers every characteristic with an equal weight during the k-means procedure (after an initial standardization step) and thus every characteristic is arguably equally important when calculating distances or sample variances. We can nevertheless note that, by construction, the k-means procedure (Algorithm 1) assigns points to the closest (in Euclidean distance) cluster centroids, and so it is cluster centroids that entirely determine the allocation of points (i.e. firms) to clusters. For a single dimension (i.e. coordinate of a point, or characteristic of a firm), we can reason that if cluster centroids are far apart (in Euclidean distance) from one another, then we may interpret this dimension as being important in partitioning points into clusters according to some notion of similarity. Conversely, if cluster centroids are close together in this dimension, we may interpret it as being less important.

By exploiting the relationship between Euclidean distance and sample variance[21] and the comparable units between each dimension,[22] we can calculate the sample variance of the cluster centroids along a single dimension (i.e. for a single coordinate of the cluster centroids) and compare these sample variances between dimensions (i.e. characteristics). This enables us to rank each dimension (i.e. characteristic) according to the sample variances of the cluster centroids along that dimension.

The results of such an exercise are presented in Table 17. Here, our notion of a characteristic's "importance" to a firm's cluster membership is based on the argument above. One immediate observation is that there is stability across slices in the rankings of the 20 most "important" characteristics. One conclusion is that the most "important" firm-level characteristics – such as *sfe*, *operprof*, *pchsale_pchinvt*, *chpmia*, *sue* & *egr* – are fundamental or analyst-based, not derived from previous returns. Another conclusion is that size (*mve*) is not one of the most "important" variables according to our clustering outcomes, in contrast to what Patton and Weller (2019) found when clustering in the cross-section using a smaller set of characteristics.

[21] It can be shown that a set of observations with a higher sample variance than another set of observations also has a higher sum of squared differences (i.e. squared Euclidean distances) between its points than the second set does.

[22] A prerequisite to apply k-means is standardizing each of the input dimensions, which we have done.

Table 15: Fraction (%) of firms in a given cluster that belong to a given industry.

Slice	Cluster	agriculture	construction	finance	manuf-acturing	mining	noclassif	retail	services	transport & utilities	wholesale
1	1	0.31	1.25	8.68	56.68	5.24	0.23	6.96	11.96	5.08	3.60
1	2	0.17	0.96	5.95	42.64	2.94	0.06	7.53	29.39	6.80	3.57
1	3	0.79	1.19	12.70	32.14	3.57	0.00	10.71	24.60	11.90	2.38
2	1	0.33	1.32	8.61	57.62	5.05	0.25	6.71	11.42	5.05	3.64
2	2	0.22	1.04	6.24	42.78	2.79	0.05	7.88	28.50	7.11	3.39
2	3	0.00	1.13	12.03	39.47	4.89	0.00	7.89	21.43	10.15	3.01
3	1	0.35	1.40	8.81	57.42	5.06	0.26	6.37	11.61	5.06	3.66
3	2	0.23	1.15	6.62	43.50	2.82	0.06	7.94	27.50	6.90	3.28
3	3	0.00	1.05	10.47	40.31	6.02	0.00	7.85	20.68	10.73	2.88
4	1	0.37	1.49	8.86	57.37	5.13	0.28	6.34	11.29	5.13	3.73
4	2	0.24	1.14	7.15	44.02	2.88	0.06	7.99	26.67	6.67	3.18
4	3	0.00	1.22	9.80	42.65	6.53	0.20	6.94	19.80	9.80	3.06
5	1	0.40	1.60	8.72	57.41	5.31	0.30	6.31	11.12	5.01	3.81
5	2	0.19	1.12	7.75	45.12	3.12	0.06	7.62	25.31	6.56	3.12
5	3	0.00	1.11	8.72	42.86	6.31	0.19	6.31	21.71	10.02	2.78
6	1	0.43	1.60	8.40	57.13	5.53	0.32	6.38	11.06	5.32	3.83
6	2	0.21	1.24	8.01	45.86	3.18	0.07	7.80	24.38	6.35	2.90
6	3	0.00	1.33	8.29	43.12	6.80	0.00	5.80	21.89	10.12	2.65
6	4	0.00	0.00	0.00	42.42	0.00	0.00	0.00	57.58	0.00	0.00

Note: Rows add up to 100%. All firms in our sample are represented.

Table 16: Fraction (%) of firms in a given industry that belong to a given cluster.

Slice	Cluster	agriculture	construction	finance	manuf-acturing	mining	noclassif	retail	services	transport & utilities	wholesale
1	1	44.44	44.44	44.76	46.50	52.34	75.00	35.74	20.84	30.23	40.00
1	2	33.33	47.22	42.34	48.30	40.62	25.00	53.41	70.71	55.81	54.78
1	3	22.22	8.33	12.90	5.20	7.03	0.00	10.84	8.45	13.95	5.22
2	1	50.00	42.11	41.60	43.97	48.80	75.00	32.93	19.27	27.98	38.60
2	2	50.00	50.00	45.60	49.40	40.80	25.00	58.54	72.77	59.63	54.39
2	3	0.00	7.89	12.80	6.63	10.40	0.00	8.54	7.96	12.39	7.02
3	1	50.00	40.00	39.45	41.96	44.62	75.00	30.29	19.28	26.48	38.18
3	2	50.00	50.00	44.92	48.21	37.69	25.00	57.26	69.28	54.79	51.82
3	3	0.00	10.00	15.62	9.82	17.69	0.00	12.45	11.45	18.72	10.00
4	1	50.00	39.02	36.26	39.50	40.74	60.00	28.94	18.28	25.70	37.04
4	2	50.00	46.34	45.42	47.08	35.56	20.00	56.60	67.07	51.87	49.07
4	3	0.00	14.63	18.32	13.42	23.70	20.00	14.47	14.65	22.43	13.89
5	1	57.14	40.00	33.72	37.55	38.69	60.00	28.77	17.54	23.92	36.89
5	2	42.86	45.00	48.06	47.31	36.50	20.00	55.71	63.98	50.24	48.54
5	3	0.00	15.00	18.22	15.14	24.82	20.00	15.53	18.48	25.84	14.56
6	1	57.14	36.59	32.24	36.41	37.41	75.00	28.85	17.11	24.63	38.30
6	2	42.86	43.90	47.35	45.02	33.09	25.00	54.33	58.06	45.32	44.68
6	3	0.00	19.51	20.41	17.63	29.50	0.00	16.83	21.71	30.05	17.02
6	4	0.00	0.00	0.00	0.95	0.00	0.00	0.00	3.12	0.00	0.00

Note: Columns add up to 100% within a slice. All firms in our sample are represented.

Table 17: Ranks of characteristics' importance during the cluster formation process, as measured by dispersion of cluster centroids.

Rank	Slice 1	Slice 2	Slice 3	Slice 4	Slice 5	Slice 6	Aggregate
1	age	ipo	ipo	ipo	ipo	sin	ipo
2	ipo	age	operprof	age	age	age	age
3	sfe	sfe	age	pchsale_pchinvt	sfe	ipo	sfe
4	securedind	operprof	sfe	sfe	securedind	sfe	sin
5	pctacc	cinvest	pchsale_pchinvt	operprof	chpmia	chpmia	securedind
6	chfeps	sue	egr	securedind	sue	securedind	operprof
7	sue	securedind	securedind	chpmia	operprof	pchsale_pchinvt	pchsale_pchinvt
8	roaq	egr	chtx	pchsale_pchrect	pchsale_pchrect	operprof	chpmia
9	pchsale_pchrect	pchsale_pchinvt	bm	egr	pchsale_pchinvt	acc	sue
10	chtx	bm	bm_ia	acc	roeq	ms	egr
11	cfp	bm_ia	cashpr	cfp_ia	acc	cfp_ia	pchsale_pchrect
12	cfp_ia	cfp	pchgm_pchsale	cfp	indmom	cfp	cfp_ia
13	grltnoa	cfp_ia	cfp_ia	rd	cfp_ia	cash	cfp
14	chatoia	chfeps	cfp	sue	cfp	roeq	cinvest
15	bm	rd	pctacc	pctacc	rd	rd	pctacc
16	bm_ia	chpmia	sue	ms	pchgm_pchsale	sue	chtx
17	chpmia	cashdebt	pchsale_pchrect	chtx	ms	pctacc	rd
18	rd	roaq	rd	cinvest	pctacc	cashpr	bm
19	ms	rsup	acc	pchsale_pchxsga	baspread	convind	acc
20	cashpr	chtx	cinvest	convind	convind	baspread	bm_ia

Note: Only the top 20 ranks are shown. The ranking procedure is based on sample variances of cluster centroid coordinates (as discussed in the text). A characteristic's aggregate ranking is based on the mean of its individual scores after they have been normalized by per-slice totals. Clusters based on the full sample of firms are considered.

C Regularizing linear models

C.1 Introduction

The predictive models we use in this study take the form of a linear regression model in d dimensions,

$$y = w'x, \tag{14}$$

where w, x are d-dimensional vectors and y is a scalar. For simplicity we do not include explicit intercept or noise terms in this formulation. Take n samples available on which to estimate such a model, and recall that there are d variables/dimensions in each sample. We stack the samples together into an $n \times d$ data matrix $\mathbf{X}$ and $n \times 1$ vector $\mathbf{y}$. Our objective is to estimate a weights vector $\mathbf{w}$ so that the linear regression model (14) holds for all samples:

$$\mathbf{y} = \mathbf{w}'\mathbf{X}. \tag{15}$$

To achieve this, one might wish to estimate the model using the OLS procedure. This would involve optimizing the weights vector $\mathbf{w}$ to minimize the residual sum-of-squares $\text{RSS}(\mathbf{w})$, which would lead to the well-known closed-form solution

$$\begin{aligned} \widehat{\mathbf{w}}_{\text{OLS}} &= \arg\min_{\mathbf{w}} \text{RSS}(\mathbf{w}) \\ &= \arg\min_{\mathbf{w}} \sum_{m=1}^{n} (y_m - \mathbf{w}'\mathbf{x}_m)^2 \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \end{aligned} \tag{16}$$

This is valid as long as $\mathbf{X}$ is full-rank, so that $\mathbf{X}'\mathbf{X}$ is invertible; otherwise, the OLS estimation problem is *ill-posed*. When the input samples are high-dimensional this is often the case and OLS cannot be used. Our study involves high-dimensional prediction.

Tikhonov and Arsenin (1977) introduced the concept of *regularization* to solve such ill-posed estimation problems. The particular form of regularization that we employ in this study is to penalize the weights vector $\mathbf{w}$ during the estimation procedure. More precisely, we compute some norm $||\mathbf{w}||$ of the weights vector and add it to the objective function that we wish to minimize, while weighting the relative degree of penalization using a hyperparameter $\lambda > 0$:

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \text{RSS}(\mathbf{w}) + \lambda||\mathbf{w}||. \tag{17}$$

Notice that our regularized/penalized optimization problem (17) augments the classical OLS optimization problem (16) with a penalty term $\lambda||\mathbf{w}||$. This has the consequence of the optimization procedure producing an estimate $\widehat{\mathbf{w}}$ that has a lower norm $||\widehat{\mathbf{w}}||$ than it otherwise would have if no penalization were applied.

Another point worth noting is that we must tune (i.e., pick an optimal value for) the λ hyperparameter. Since our data involve time dependencies, we use the slicing procedure in Section 3.2 to tune all hyperparameters out-of-sample in a way that respects the temporal dependencies.

Finally, note that the linear regression problem will retain its original forms (14) and (15). This is because the regularization procedure results in a (more suitable) estimate of the weights w using the available samples while not affecting the linear functional form of the regression model. This has the advantage of allowing us to easily introduce ML techniques into the models of firm-level heterogeneity that we described in Section 3.1.

C.2 Regularized methods used in this study

We now make the estimation problem (17) concrete.

Ridge regression

If we use the square of the ℓ_2 norm, $||\mathbf{w}||_2^2 = \mathbf{w}'\mathbf{w}$, as our penalization term, we obtain another closed-form solution,

$$\begin{aligned}\widehat{\mathbf{w}}_{\text{Ridge}} &= \arg\min_{\mathbf{w}} \text{RSS}(\mathbf{w}) + \lambda||\mathbf{w}||_2^2 \\ &= (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{y},\end{aligned}$$

and this technique is called *Ridge regression*. It is due to Hoerl and Kennard (1970).

Note that $\mathbf{I}$ is the identity matrix, so the closed-form expression above effectively adds some weight λ to the diagonals of $\mathbf{X}'\mathbf{X}$ before inverting it. This illustrates how the potentially ill-conditioned term $\mathbf{X}'\mathbf{X}$ is made invertible. It also implies that the elements of $\widehat{\mathbf{w}}_{\text{Ridge}}$ are shrunk towards zero, with the degree of shrinkage increasing in λ.

Lasso

If we use the ℓ_1 norm, $||\mathbf{w}||_1 = \sum_{m=1}^{d} |w_m|$, as our penalization term,

$$\widehat{\mathbf{w}}_{\text{Lasso}} = \arg\min_{\mathbf{w}} \text{RSS}(\mathbf{w}) + \lambda||\mathbf{w}||_1$$

then this technique is called the *Lasso*. It is due to Tibshirani (1996).

As we explained in Section 2.2, the estimated coefficient vector $\widehat{\mathbf{w}}_{\text{Lasso}}$ will tend to be *sparse*; that is, to have zero elements in place of elements with a small magnitude. Wainwright (2009) and Tropp (2006) explain that the Lasso can be interpreted in terms of variable selection: the intuition is that the ℓ_1 penalty is the closest convex relaxation of the ℓ_0 discrete variable selection penalty. This property means that predictive variables whose coefficients are non-zero can be directly interpreted as Lasso-selected variables.

ElasticNet

If we use a convex combination of the ℓ_1 and squared ℓ_2 norm elements as the penalization term,

$$\widehat{\mathbf{w}}_{\text{ElasticNet}} = \arg\min_{\mathbf{w}} \text{RSS}(\mathbf{w}) + \lambda \sum_{m=1}^{d} \left[\alpha w_m^2 + (1-\alpha)|w_m|\right]$$

then this technique is called the *ElasticNet*. It is due to Zou and Hastie (2005). Although often used in practice, one disadvantage of the ElasticNet for our purposes is that it requires an additional hyperparameter $\alpha \in (0, 1)$ to be tuned.

D Other methods

We preview some additional machine learning techniques that could be applied to the heterogeneous predictability problem.

D.1 Multitask learning

Multi-task learning models are supervised learning models comprised of individual "task" models that are estimated jointly. Here, "tasks" are best thought of as different problem domains that are nevertheless related to one another. Argyriou et al. (2006) formulate the problem in such a way that individual task models "borrow strength" from one another through shrinkage of their coefficients to pooled values. Variations – including the possibility of task clustering – are discussed in Evgeniou and Pontil (2004). Other extensions, such as by Jalali et al. (2010), modify the forms of shrinkage imposed.

We will now formulate the original multitask learning paper of Argyriou et al. (2006) in terms of firms and characteristics.

Assume I groups of firms (indexed by $i \in \{1, 2, \ldots, I\}$), with each group consisting of some J firms (with J varying from group to group), represented by row vectors $\mathbf{x}_{ij}, j \in \{1, 2, \ldots, J\}$. Assume p partworths, i.e., each vector $\mathbf{x}_{ij}$ has p columns, and these may represent firm- or market-level characteristics. Define $\mathbf{X}_i$ as the $J \times p$ design matrix for group i (each row of this matrix corresponds to one firm); by $\mathbf{w}_i$ the $p \times 1$ column vector of the partworths for group i; and by $\mathbf{Y}_i$ the $J \times 1$ column vector containing the returns of the firms in group i.

In the above notation, the prediction task is therefore

$$\mathbf{Y}_i = \mathbf{x}_{ij}\mathbf{w}_i.$$

The estimation procedure is

$$\begin{aligned} \min_{\{\mathbf{w}_i\},\mathbf{w}_0,D} \quad & \tfrac{1}{\gamma}\sum_{i=1}^{I}\sum_{j=1}^{J}\left(y_{ij} - \mathbf{x}_{ij}\mathbf{w}_i\right)^2 + \sum_{i=1}^{I}(\mathbf{w}_i - \mathbf{w}_0)' D^{-1} (\mathbf{w}_i - \mathbf{w}_0) \\ \text{subject to} \quad & D \text{ is a positive semidefinite matrix scaled to have trace 1} \end{aligned}$$

This shrinks partworths $\mathbf{w}_i$ towards a common vector $\mathbf{w}_0$. It also allows heterogeneity between the groups i, while explicitly pooling information across the groups

Argyriou et al. (2006) derive a dual formulation for the estimation procedure above that is convex in each variable, and so may be estimated using an iterative procedure. Other relevant multi-task learning models are proposed by Evgeniou and Pontil (2004) and Jalali et al. (2010).

D.2 Group Lasso

The *Group Lasso* model of Yuan and Lin (2006), and its extensions, allow regularization/penalization of groups of coefficients. Freyberger et al. (2020) made use of the Group Lasso to select/shrink blocks of characteristic-related coefficients. The approach could also be applied to select/shrink blocks of industry-related coefficients or other groupings of coefficients that correspond to some heterogeneous predictability setting.

Returning to the original formulation (15), Yuan and Lin (2006) have established a variant of the Lasso known as the *Group Lasso* that makes use of a grouping structure between predictive variables. If there are J such groups of predictive variables, then we divide our coefficient vector $\mathbf{w}$ into J groups and denote them $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_J$. The Group Lasso estimation procedure makes use of J individual ℓ_2

norms of these subsets and solves

$$\widehat{\mathbf{w}}_{\text{Group Lasso}} = \arg\min_{\mathbf{w}} \text{RSS}(\mathbf{w}) + \lambda \sum_{j=1}^{J} ||\mathbf{w_j}||_2.$$

Given an appropriate grouping structure $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_J$, the Group Lasso (and its variants) may also be useful building blocks for formulating heterogeneous prediction models.