Internet Appendix for "Uncovering sparsity and heterogeneity in firm-level return predictability using machine learning"

Theodoros Evgeniou, Ahmed Guecioueur, Rodolfo Prieto*

Contents

*INSEAD, Bd de Constance, 77300 Fontainebleau, France, e-mail: theodoros.evgeniou@insead.edu, ahmed.guecioueur@insead.edu, and rodolfo.prieto@insead.edu.

A Characteristics

Tables IA.1 and IA.2 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 IA.1: Market-level variables. 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 IA.2: Firm-level characteristics. 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

Code	Name	Frequency	Type	Description
chcsho	% change in shares outstanding	Yearly	percentage	Annual percentage change in shares outstanding (csho)
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

Code	Name	Frequency	Type	Description
egr	% Change in common shareholder equity	Yearly	percentage	Annual percent change in book value of equity (ceq)
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

Code	Name	Frequency	Type	Description
mom1m	1-month-momentum	Monthly	percent	1-month-momentum
mom36m	36-month-momentum	Monthly	percent	Cumulative returns from months t -36 to t - 13
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

Code	Name	Frequency	Type	Description
pchsaleinv	% Change in Sales-to-Inventory	Yearly	percentage	% change in (sales/inventory)
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

Code	Name	Frequency	Type	Description
sfe	Analyst mean annual earnings forecast (scaled)	Quarterly	ratio	Analyst mean annual earnings forecast scaled by absolute price per share at fiscal quarter end
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 cluster formation, beginning with the k-means algorithm itself.

B.1 k-means algorithm

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

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: $\gamma_i = \arg\min_k ||x_i - \mu_k||^2$
4: Update the means of the currently assigned clusters: $\mu_k = \frac{1}{N_k} \sum_{i:\gamma_i=k} x_i$
5: **until** convergence

B.2 Firm distributions across clusters

As a first indication of how the inferred clusters of firms vary from one slice to another, Table IA.3 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.

Figure 1 in the paper indicates that clusters are stable in principal component space. We now supplement this visual evidence on cluster stability by counting the fraction of firms that remain in the same cluster from slice to slice, with results shown in Table IA.4. We find that firms' cluster memberships are highly stable once firms enter the two largest clusters (1 & 2). The third, smaller cluster shows more variation as a number of firms in our sample leave this

[1] 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 (2022).

cluster to join others. This can occur as they age or improve in profitability; nevertheless, the 84% level of persistence observed for this third cluster is also high.

Table IA.3: Cluster counts. 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

Table IA.4: Firm-cluster stability. Cells denote the fraction of firms (%) remaining in the same cluster from one slice to the next.

Cluster	Slice 1 to 2	Slice 2 to 3	Slice 3 to 4	Slice 4 to 5	Slice 5 to 6	Unconditional
1	97.66	99.39	98.98	99.20	98.12	98.66
2	100.00	99.94	100.00	99.93	96.45	99.31
3	40.51	89.41	88.89	82.96	99.61	84.16

B.3 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 IA.5 and IA.6 characterize firms' joint membership of industries and clusters. It is clear from the joint memberships that clusters do not span industries.

We now reinforce this point by considering individual characteristics one at a time (rather than cluster membership as a summary) and relating these to a firm's industry membership. To do so, we compute industry means of firm-level characteristics on the training set of the largest slice (i.e. slice 6, in order to facilitate a comparison with the cluster formation process), and then take the standard deviation of each across all industries. Because all characteristics are already scaled cross-sectionally to the same range of $[-1, +1]$ per month, as we describe in the paper, the across-industry standard deviations will be in the same units, thus allowing us to concisely summarize which characteristics vary the most across industries. We sort characteristics from greatest across-industry variation to least, and list the 20 highest-variation characteristics in Table IA.7.

It is instructive to compare characteristics' variation across industries, in Table IA.7, with characteristics' variation across clusters, which we plot in Figure 2 of the paper. There are some similarities in which characteristics vary the most across industries and clusters, but

also important differences, which is consistent with our previously discussed findings that cluster and industry memberships are distinct (according to Tables IA.5 and IA.6). Interpreting Table IA.7, the *age* characteristic exhibits substantial variation across industries, but firm size (*mve*) varies even more. This is a major difference to cluster formation, for which firm size is *not* one of the characteristics that best explains cluster differences. On a related note, IPO status, which is another age-related indicator, does not even appear in the top 20 highest variation characteristics across industries. Similarly, a number of variables related to analyst forecasts (*sfe*, *sue*), sales growth (*pchsale_pchinvt*) and other accounting/fundamental variables (*acc*, *operprof*, *roeq*, for example) also do not make the top 20 most variable characteristics across industries, but do appear in the list of distinctive cluster-specific characteristics, as detailed in Figure 2 in the paper. Conversely, a number of industry-specific characteristics (*herf*, *indmom*), fundamental characteristics (*tang*, *ps*) and price-based characteristics (*baspread*, *beta*) do appear in Table IA.7, but are not characteristics that explain cluster differences well.

Table IA.5: Cluster vs. industry counts. 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 IA.6: Industry vs. cluster counts. 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 IA.7: Characteristic variation across industries. Firm-level characteristics ranked by their importance to industry definitions, as measured by their (comparable in units) across-industry standard deviations.

Rank	Characteristic	SD of Industry Means
1	securedind	0.35
2	indmom	0.35
3	mve	0.23
4	rd	0.20
5	age	0.19
6	ms	0.16
7	cash	0.14
8	dolvol	0.12
9	tang	0.11
10	ps	0.11
11	convind	0.11
12	herf	0.10
13	nanalyst	0.10
14	baspread	0.09
15	orgcap	0.06
16	betasq	0.06
17	divo	0.06
18	mom12m	0.05
19	idiovol	0.05
20	beta	0.05

B.4 Importance of characteristics in cluster formation

We have provided an interpretation of the clusters in terms of their compositions in Section 5.1 of the paper. 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 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[2] and the comparable units between each dimension,[3] 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 IA.8. Here, our notion of a characteristic's importance to a firm's cluster membership is based on the argument above. It is immediately evident that there is stability across slices in the rankings of the 20 most important characteristics. One observation 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 observation is that size (*mve*) is not one of the most "important" variables according to our clustering outcomes, in contrast to what Patton and Weller (2022) found when clustering in the cross-section using a smaller set of characteristics.

[2]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.

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

Table IA.8: Interpreting clusters using dispersions. 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 Alternative text-based industry partition

In our paper, we defined our baseline industry partition based on SIC code ranges. We now consider an alternative specification, applying our aggregate out-of-sample predictability analysis using a definition of industry membership based on Hoberg and Phillips (2016)'s Text-Based Network Industry Classifications (TNIC).[4]

There are a few hurdles to implementing this analysis, so we briefly describe them and our solutions. The first hurdle is that it is unclear how many TNIC-based industries should be used as a partition, given the absence of a principled method to do so by Hoberg and Phillips (2016) themselves. We therefore define 5 TNIC-based industries as a compromise between the 3-4 characteristics-based clusters and the 10 SIC-based industries in our existing analyses.

The second hurdle is that the Hoberg and Phillips (2016) data library makes available only a set of 25 industries formed with respect to their first full year of data (1997), while we wish to use a more recent snapshot to ensure classifications are as up-to-date as possible. We also wish to use a partition comparable in granularity to our existing partitions (i.e. 5 industries is comparable to our SIC and clusters membership). We therefore recompute a set of 5 TNIC-based industries using firm-pairwise similarity scores from 2009, following a procedure similar to that laid out by Hoberg and Phillips (2016, Appendix B). The procedure involves recursively merging industries that are closest (i.e. that have the highest *industry*-pairwise mean similarity scores) and repeating the procedure to achieve the desired number of industries. As Hoberg and Phillips (2016, Appendix B) point out, this approach is somewhat conservative because there may exist other partitions with the same within-industry similarity scores; therefore, if exploiting this particular (non-unique) TNIC partition using our procedure enables us to measure some out-of-sample predictability, this does not rule out the possibility that even more substantial levels of predictability could be detected using other variants of the Hoberg and Phillips (2016) TNIC.

Once we have uniquely assigned each firm in our partition to one of the 5 TNIC-based industries, we estimate the usual set of models on this new industry partition of the cross-section, using the usual validation & test sets. The resulting R^2_{OOS} values are displayed in Table IA.9.

Comparing Panel (c) of Table IA.9 to Panel (c) of Tables 4 & 5 in the paper, the highest R^2_{OOS} of 0.95% on the full sample of firms lies in between similar values using other partitions of the cross-section (0.76% on SIC-based industry groupings, and 1.05% on cluster groupings). The magnitude remains lower than that attained by using the cluster groupings (in Table 5 of the paper), and, importantly, is achieved by a predictive model that incorporates both

[4]For a description of the use of various industry groups in the study of stock returns see e.g., Chan et al. (2007)

heterogeneity and sparsity (By-industry Lasso).

This exercise confirms that exploiting the alternative industry partition derived from Hoberg and Phillips (2016)'s TNIC can also produce more accurate return forecasts than when this information is ignored.

Table IA.9: Predictability using TNIC industries. Aggregate out-of-sample predictability, measured by R^2_{OOS} (%), when partitioning firms by industry membership according to Hoberg and Phillips (2016)'s Text-Based Network Industry Classifications (TNIC).

Panel (a)

Model	Top 1,000
Two-stage Ridge	1.75
Pooled ElasticNet	1.74
Pooled Ridge	1.74
By-industry ElasticNet	1.72
By-industry Ridge	1.66
Pooled Lasso	1.66
By-industry Lasso	1.57
Two-stage Lasso	1.53
Pooled OLS	-9.66
By-industry OLS	-13.21
Two-stage OLS	-13.21

Panel (b)

Model	Top 2,000
Two-stage Ridge	1.53
Pooled ElasticNet	1.50
Pooled Ridge	1.50
By-industry Ridge	1.47
Pooled Lasso	1.46
By-industry Lasso	1.40
Two-stage Lasso	1.36
By-industry ElasticNet	1.30
Pooled OLS	-8.98
By-industry OLS	-12.16
Two-stage OLS	-12.16

Panel (c)

Model	All Firms
By-industry Lasso	0.95
Pooled Lasso	0.94
Two-stage Ridge	0.92
Pooled Ridge	0.91
By-industry ElasticNet	0.90
By-industry Ridge	0.90
Pooled ElasticNet	0.90
Two-stage Lasso	0.87
Pooled OLS	-5.49
By-industry OLS	-7.20
Two-stage OLS	-7.20

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

D Small firms

We explore our ability to detect out-of-sample predictability on relatively smaller firms by re-estimating the predictive models on the subset of the smallest 1,000 and smallest 2,000 firms by market capitalization, keeping the same industry and cluster groupings as our main results on overall predictability.

These results are shown in Panels (a) and (b), respectively, of Tables IA.10 and IA.11. We continue to show the main results on the full set of firms in Panel (c) of each table, which are therefore identical to Panel (c) of Tables 4 and 5 in the paper.

A few conclusions can be drawn from Tables IA.10 and IA.11. First, recalling that models estimated on the largest firms detected a higher level of out-of-sample predictability than models estimated on the full set of firms in our paper, it is therefore not surprising that the level of predictability deteriorates when models are estimated on the smallest firms, though

it remains high: for example, the By-cluster Lasso model detects an R^2_{OOS} value of 0.70% when estimated on the smallest 1,000 firms, and an R^2_{OOS} value of 0.98% when estimated on the smallest 2,000 firms, in Table IA.11 Panels (a) and (b), respectively. Second, using cluster partitions rather than industry partitions allows us to detect higher levels of OOS predictability when we condition on the smallest firms when estimating the models, as can be seen by comparing the general R^2_{OOS} magnitudes between Tables IA.10 and IA.11. Third, the best-performing model when training on the smallest 1,000 firms is the By-cluster Lasso, which is a sparse model, so the cluster partition continues to demonstrate its usefulness for detecting a parsimonious set of predictive characteristics for the smallest firms in the cross-section. Therefore, this analysis confirms a number of our main findings in the paper when applied to the smallest firms in our sample.

We can also draw a high-level contrast to the results reported by GKX, with the usual caveats that our time periods and samples differ. GKX reported their highest R^2_{OOS} of 0.47% (attained by a deep neural net) on the bottom 1,000 firms by size. Our results in Table IA.11 Panel (a) show our highest R^2_{OOS} on our bottom 1,000 firms by size is 0.70%, attained by the By-cluster Lasso. Therefore, our results remain higher in magnitude on this particular subset of firms.

We now examine the frequency of selection of coefficients for this By-cluster Lasso model estimated on the smallest 1,000 firms, with results shown in Table IA.12. These results indicate that a more diverse, larger set of characteristics can be selected when estimating on the smallest stocks only: 15 variables appear in Table IA.12. As a comparison, when we estimated the same predictive model on all available firms, 5 predictive variables were selected (in Table 7 of the paper) and these are a direct subset of the 15 selected here in Table IA.12. At the same time, only a subset of the full set of predictive variables used in our study are ever selected, so we argue that our methodology continues to show its worth for uncovering sparsity among the zoo of variables that can predict next-month firm-level returns successfully out-of-sample.

Table IA.10: Small-firm predictability using industries. Aggregate out-of-sample predictability, measured by R^2_{OOS} (%), when partitioning firms by industry membership.

Panel (a)

Model	Bottom 1,000
Pooled Lasso	0.38
Pooled ElasticNet	0.37
Two-stage Lasso	0.35
Pooled Ridge	0.23
By-industry Lasso	0.21
Two-stage Ridge	0.19
By-industry ElasticNet	0.16
By-industry Ridge	0.15
Pooled OLS	-2.99
By-industry OLS	-4.56
Two-stage OLS	-4.56

Panel (b)

Model	Bottom 2,000
Two-stage Ridge	0.53
Pooled Ridge	0.52
By-industry Ridge	0.50
Pooled ElasticNet	0.49
By-industry ElasticNet	0.46
Pooled Lasso	0.35
Two-stage Lasso	0.33
By-industry Lasso	0.26
Pooled OLS	-3.69
By-industry OLS	-5.10
Two-stage OLS	-5.10

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 firms and then generating predictions for that same subset: (a) on the smallest 1,000 firms by market capitalization, (b) on the smallest 2,000 firms, and (c) on the full sample.

Table IA.11: Small-firm predictability using clusters. Aggregate out-of-sample predictability, measured by R^2_{OOS} (%), when partitioning firms by cluster membership.

Panel (a)

Model	Bottom 1,000
By-cluster Lasso	0.70
By-cluster ElasticNet	0.70
Pooled Ridge	0.64
Two-stage Ridge	0.62
By-cluster Ridge	0.59
Two-stage Lasso	0.54
Pooled ElasticNet	0.52
Pooled Lasso	0.48
Pooled OLS	-4.43
By-cluster OLS	-32.25
Two-stage OLS	-32.25

Panel (b)

Model	Bottom 2,000
By-cluster ElasticNet	0.99
By-cluster Lasso	0.98
Two-stage Lasso	0.98
Pooled Ridge	0.94
Two-stage Ridge	0.94
Pooled ElasticNet	0.91
By-cluster Ridge	0.89
Pooled Lasso	0.89
Pooled OLS	-4.69
By-cluster OLS	-52.93
Two-stage OLS	-52.93

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 firms and then generating predictions for that same subset: (a) on the smallest 1,000 firms by market capitalization, (b) on the smallest 2,000 firms, and (c) on the full sample.

Table IA.12: Small-firm By-cluster Lasso selected predictors. Frequency of selection (% of slices) of characteristics by cluster, when estimating the by-cluster Lasso model on the smallest 1,000 firms by market cap.

Characteristic	Cluster 1	Cluster 2	Cluster 3	Cluster 4
(Intercept)	100	100	100	100
baspread	0	0	33	100
cashpr	33	17	17	0
chfeps	0	17	0	0
chmom	0	0	17	0
chpmia	0	0	17	100
dfy_mkt	0	17	0	0
dp_mkt	33	17	17	0
ep_mkt	17	33	0	0
lev	0	17	0	0
mom36m	0	17	17	0
mve	0	17	0	0
pchgm_pchsale	0	0	0	100
pctacc	0	17	0	0
salecash	0	17	0	0
sue	33	17	33	100

Note: The model was estimated based on the smallest 1,000 firms, once per slice.

E Regularizing linear models

E.1 Introduction

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

$$y = w'x, \tag{1}$$

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 (1) holds for all samples:

$$\mathbf{y} = \mathbf{w}'\mathbf{X}. \tag{2}$$

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 RSS($\mathbf{w}$),

which would lead to the well-known closed-form solution

$$\begin{aligned}\widehat{\mathbf{w}}_{\text{OLS}} &= \arg\min_{\mathbf{w}} \text{RSS}(\mathbf{w}) \\ &= \arg\min_{\mathbf{w}} \sum_{m=1}^{n} (y_m - \mathbf{w}'\mathbf{x}_m)^2 \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.\end{aligned} \tag{3}$$

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 $\lambda > 0$:

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \text{RSS}(\mathbf{w}) + \lambda||\mathbf{w}||. \tag{4}$$

Notice that our regularized/penalized optimization problem (4) augments the classical OLS optimization problem (3) 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 of the paper 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 (1) and (2). This is because the regularization procedure results in a (more suitable) estimate of the weights $\mathbf{w}$ 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 of the paper.

E.2 Regularized methods used in this study

We now make the estimation problem (4) 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 of the paper, 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,[5] one disadvantage of the ElasticNet for our purposes is that it requires an

[5]A more general machine learning technique is Multitask Learning, see e.g., Argyriou et al. (2006), Evgeniou and Pontil (2004), and Jalali et al. (2010).

additional hyperparameter $\alpha \in (0, 1)$ to be tuned.

F Bibliography

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

Chan, L.K., Lakonishok, J., Swaminathan, B., 2007. Industry classifications and return comovement. Financial Analysts Journal 63, 56–70.

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.

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.

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.

Hoberg, G., Phillips, G., 2016. Text-based network industries and endogenous product differentiation. Journal of Political Economy 124, 1423–1465.

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

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.

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. pp. 281–297.

Patton, A.J., Weller, B., 2022. Risk price variation: The missing half of empirical asset pricing. Review of Financial Studies (forthcoming).

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

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.

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.