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

Theodoros Evgeniou Ahmed Guecioueur Rodolfo Prieto

INSEAD

Contributions

For the problem of firm-level month-ahead return prediction, and interpreting characteristic importance,

- We find statistical evidence (using the bootstrap) that heterogeneity matters for predictability.
- By incorporating heterogeneity in predictive models, we improve their out-of-sample performance.
- We highlight new perspectives on characteristics:
  - Different characteristics can matter for different groups of firms.
  - Characteristics can be used to infer firm groupings, in addition to directly predicting returns.
- We uncover sparsity in the cross-section using lasso-based models, without sacrificing predictability.

Incorporating heterogeneity in linear predictive models

- Index firms by i, and let c_{it} be a high-dimensional (M-dim.) vector of a firm's characteristics.
- We apply ML regularization techniques to classic *pooled* linear models with common coefficients:

$$r_{i,t+1} = \alpha + \theta_1 c_{it1} + \theta_2 c_{it2} + \ldots + \theta_M c_{itM} \quad (1)$$

$$= \alpha + \theta' c_{it} \quad (2)$$

- Furthermore, we incorporate heterogeneity in predictive relationships in *by-group* models, given a mapping from a firm i to its (unique) group j, by employing group-specific coefficients:

$$r_{i,t+1} = \alpha_j + \theta_{1j} c_{it1} + \theta_{2j} c_{it2} + \ldots + \theta_{Mj} c_{itM} \quad (3)$$

$$= \alpha_j + \theta_j' c_{it} \quad (4)$$

- We also combine the two stages to specify composite *two-stage* models, that take the form

$$r_{i,t+1} = \alpha_0 + \theta_0' c_{it} + \sum_{j=1}^{K} \mathbb{I}_{i\in j}(\alpha_j + \theta_j' c_{it}) : \quad (5)$$

  1. estimate a pooled model on the entire cross-section of returns, then
  2. estimate a by-group model on the residuals of the first-stage pooled model.

- NB. need to tune multiple regularization parameters (e.g. lasso λ) for by-group and two-stage models.

Motivations for predictive heterogeneity

- Equilibrium asset pricing models with multiple state variables, such as Menzly, Santos, and Veronesi (2004) and Koijen and Yogo (2019), imply heterogeneity in firm-specific predictive relationships.
- Patton and Weller (2019) find evidence for risk premia deviations that are specific to groups of firms (rather than the whole cross-section) in a modified conditional CAPM.

Data & evaluation

- 109 predictive characteristics: 101 are firm-specific (Green, Hand, and Zhang 2017) and 8 are market-level (Welch and Goyal 2007).
- Time period: 1980-2015 (inclusive).
- Our out-of-sample evaluation uses the same R^2 metric and takes the same expanding window approach as Gu, Kelly, and Xiu (2020).

Specifications of predictive heterogeneity

To define groupings of firms, we consider two alternatives:

1. Firm industry memberships – based on SIC codes.
2. Inferring (possibly) latent group memberships from observable characteristics – by applying k-means clustering to characteristic means.

Bootstrap-based evidence in favour of incorporating heterogeneity

- For each of the 2 specifications of heterogeneity in the cross-section, we estimate pooled and by-group models and compute the incremental (by-group minus pooled) out-of-sample R^2 (%).
- Repeat for 1000 nonparametric bootstrap samples of the whole cross-section, in order to compute bootstrap confidence intervals for statistical testing.
- The statistically-significant incremental R^2 values are typically positive:

Table 1: Using industry memberships to define heterogeneity in predictive relationships.

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: Asterisks denote the significance level from two-tailed tests (***=99%, **=95%, *=90%).

Table 2: Clustering firms to define heterogeneity in predictive relationships.

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: see previous table.

Stable & interpretable clusters of firms in the cross-section

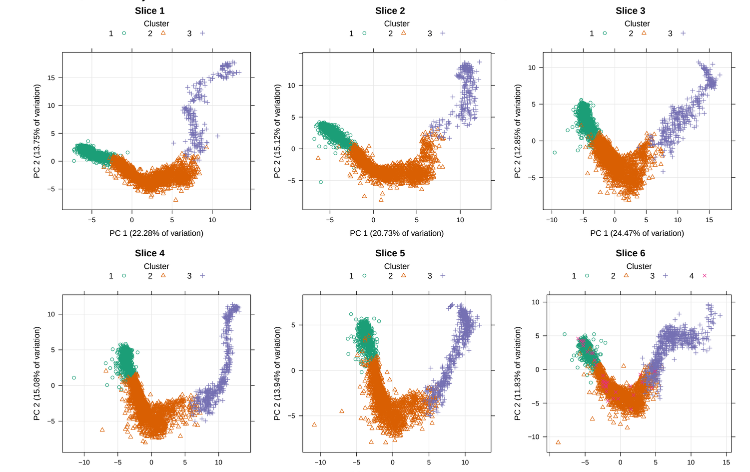

Figure 1: Cluster visualisations (in principal component space) over time. Each point represents a firm.

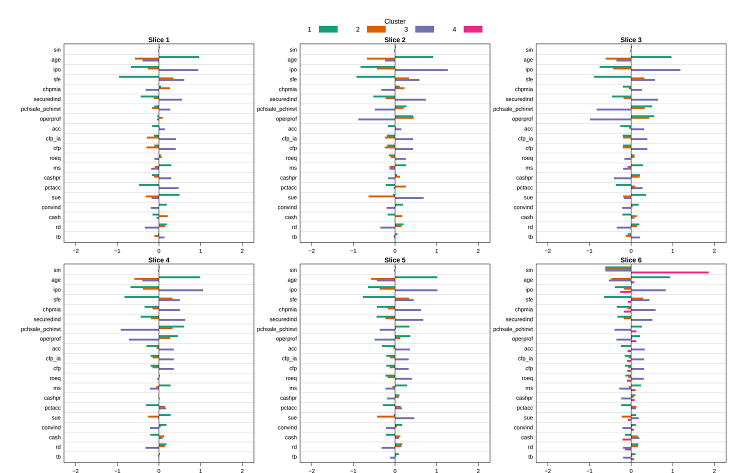

Figure 2: Interpreting clusters by comparing their characteristic means.

The 3 stable clusters are interpretable based on the characteristics of the firms they comprise:

1. Mature firms with lower *sfe* (scaled analyst earnings forecasts), lower likelihood of *securedind* (secured debt), and higher *operprof* (operating profitability) and *pchsale_pchinvt* (difference between %ge changes in sales and inventory).
2. Younger firms. Below average *cfp* (cashflow to price) and *sue* (directional earnings surprises).
3. Younger and recently IPO'ed and with low *pchsale_pchinvt* & *operprof* and high chance of *securedind*.

Overall predictability, measured by out-of-sample R^2 (%)

- Clustering firms to define heterogeneity in predictive relationships, then incorporating this specification of heterogeneity when estimating a regularized linear model achieves an out-of-sample $R^2 = 1.05\%$.
- For context, Gu, Kelly, and Xiu (2020) achieved an out-of-sample $R^2 = 0.40\%$ on their sample using a deep neural network. Rapach and Zhou (2013) argue that an in-sample $R^2 \approx 1\%$ is enough to falsify some asset pricing models, and that out-of-sample values are even lower.

Table 3: Using industry memberships to define heterogeneity in predictive relationships.

Panel (a)

Model	Top 1000
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 2000
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

Table 4: Clustering firms to define heterogeneity in predictive relationships.

Panel (a)

Model	Top 1000
Pooled Ridge	1.91
Two-stage 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 2000
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 ElasticNet	1.03
By-cluster Lasso	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

Uncovering sparsity & heterogeneity in characteristic importance

- In the overall predictability results (above), heterogeneous lasso-based linear models performed well.
- Selection by the lasso is a measure of variable importance. These lasso-selected predictive variables vary between clusters of firms, and are a sparse subset of the 109 total variables employed.
- In contrast to Gu, Kelly, and Xiu (2020), the important predictive variables are mostly a subset of low-frequency cash and profitability-related coefficients (*chpmia*, *cashpr*) and the market-level D/P ratio (*dp_sp500*), rather than higher-frequency price-based predictors.

Table 5: Frequency of selection (%) across slices of our database, according to the by-cluster lasso model.

Characteristic	Cluster 1	Cluster 2	Cluster 3
(Intercept)	100	100	100
baspread	17	0	33
cashpr	33	17	33
chpmia	33	0	33
dp_sp500	33	17	17
sue	0	0	17

References

Green, Jeremiah, John RM Hand, and X Frank Zhang. 2017. "The characteristics that provide independent information about average US monthly stock returns". *Review of Financial Studies* 30 (12): 4389–4436.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu. 2020. "Empirical asset pricing via machine learning". *Review of Financial Studies* 33 (5): 2223–2273.

Koijen, Ralph S. J., and Motohiro Yogo. 2019. "A demand system approach to asset pricing". *Journal of Political Economy* 127 (4): 1475–1515.

Menzly, Lior, Tano Santos, and Pietro Veronesi. 2004. "Understanding predictability". *Journal of Political Economy* 112 (1): 1–47.

Patton, Andrew J, and Brian Weller. 2019. "Risk price variation: the missing half of empirical asset pricing". *Economic Research Initiatives at Duke (ERID) Working Paper*, no. 274.

Rapach, David, and Guofu Zhou. 2013. "Forecasting stock returns". In *Handbook of Economic Forecasting*, 2:328–383.

Welch, Ivo, and Amit Goyal. 2007. "A comprehensive look at the empirical performance of equity premium prediction". *Review of Financial Studies* 21 (4): 1455–1508.