Interpretable factors of firm characteristics

Yuxiao Jiao (TsinghuaU), Guofu Zhou (WashU), YingziZhu (TsinghuaU).

Email: jiaoyx.18@sem.tsinghua.edu.cn

Abstract

We propose a new approach to **construct factors from firm characteristics**.

In contrast to existing studies, each of our factors comes from the same group of statistically related firm characteristics, making its **economic interpretation possible**. **The number of groups is** not chosen ad hocly, but rather **determined by data**.

Applying our method to a set of 94 representative firm characteristics, we find that the factors chosen by our approach is not only **easy to interpret economically**, but the associated factor model **outperforms** existing models.

We also apply our approach to the recent developed and highly effective **IPCA** model of Kelly, Pruitt and Su (2019), and find that our factors not only are **well linked to the associated economic risks**, but also can **price assets no worse than the standard IPCA latent factors** that are difficult to interpret.

Motivation

Background: As surveyed by Harvey, Liu, and Zhu (2016) and Hou, Xue, and Zhang (2020), there are potentially hundreds of firm characteristics or firm-level factors that affect the expected returns in the cross-section of stocks.

Questions (Cochrane, 2011):

- How many factors do we really need?
- How to explain the cross-sectional variation of expected stock returns?

Two existing approaches to answer the above questions.

- The first approach: Principal Component Analysis (PCA) and its variations (Connor and Korajczyk, 1988), (Kelly et al., 2019).
  - Disad.: the extracted factors are linear combinations of all the existing characteristics, making them **difficult to interpret economically**.
  - E.g.: four firm characteristics (2 momentum and 2 value). PCA factors are neither momentum nor value factors.
- The second approach: Machine Learning (Feng, Giglio, and Xiu, 2020), (Freyberger, Neuhierl, and Weber, 2020), (Kozak, Nagel, and Santosh, 2020).
  - Disad.: tend to **over-identify the number of factors** that matter, as it is difficult to handle and distinguish highly correlated factors by the existing ML models.
  - E.g.: many momentum factors may be selected as independent factors by LASSO, which are highly correlated.

Our paper provides a simple approach to address the problems.

- Intuitively, our method has two steps.
  - Step 1. Divide the firm characteristics into statistically related clusters.
  - Step 2. Extract optimally a factor for each cluster.

Our method can address the problems.

- Factors are easy to interpret economically.
- Can handle highly-correlated factors and will not over-identify the number of factors.

Literature Review

Our approach is inspired by Stambaugh and Yuan (2017) who appear the first to use cluster to isolate the factors.

- We apply a clustering algorithm that is applicable to high dimensional case.
- We let data to find the best factor from the cluster instead of taking the average.

Our paper is also related to literature that clusters firm characteristics based on economic meanings (intuitive clustering, denoted as IC hereafter) (Hou, Xue, Zhang, 2015; McLean and Pontiff, 2016; Hou, Xue, Zhang, 2020; Han, He, Rapach and Zhou, 2020).

- Adv. of IC: considers economic meaning.
- Disad. of IC: ignores data information.
- Our clustering result (data-driven clustering, denoted as DC hereafter) improves IC.

Data and Methodology

Data

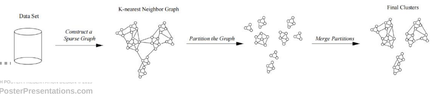
- 94 firm characteristics in Gu et al. (2020).
- Sample period: 1985.1-2021.12.
- Source: CRSP and Compustat.

Choose a Clustering Algorithm:

- Density-based algorithm: assume a probability model.
  - ×: No additional assumption.
- Spectral clustering: based on principal components.
  - ×: Sensitive to noise (Bojchevski et al., 2017) .
- Combinational algorithm:
  - √: Works directly on the observed data with no direct reference to an underlying probability model.

Clustering Algorithm: We use the cluster method called Chameleon by Karypis et al. (1999), which belongs to combinational algorithm.

- Aims: find clustering result with low inter-cluster similarity and high within-cluster similarity. OR lower IS.
  - IS: ratio of inter-cluster similarity to within-cluster similarity.
- Steps:
  - Step 1: Start from the IC. Assume similarity between firm characteristics in different clusters in IC is 0.
  - Step2. Divide firm characteristics into several sub-clusters.
  - Step3. Merge sub-clusters.



Choose the number of clusters.

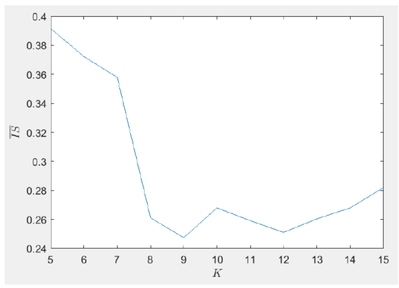
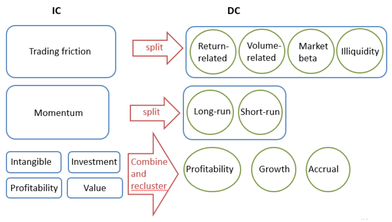


Figure 2: Average inter-cluster similarity

Results: Clustering Results

9 clusters.



Results: Performance of the DC-model

Construction of the DC-model

- Step 1. Based on firm characteristics in each cluster, measure risk exposure with predicted stock return.
  - Stock return is predicted as:

$$\hat{R}_{n,k,t+1} = a * \hat{R}^1_{n,k,t+1} + (1-a) * \hat{R}^2_{n,k,t+1}.$$

  - $\hat{R}^1_{n,k,t+1}$ is the average of the characteristics in cluster k.
  - $\hat{R}^2_{n,k,t+1}$ is the average of predicted returns from single regressions (Han et al, 2023).
  - a is weight and maximize the Sharpe ratio in the training data.
- Step 2. Estimate risk premiums.

$$R_{n,k,t+1} = \sum_{k=1}^{K} \gamma_k * \hat{R}_{n,k,t+1} + \varepsilon_{n,k,t+1}.$$

Performance of the DC-model.

- Performance of model factors (below left)

Cluster		Mean(%)	S.D.(%)	Sharpe	MDD(%)
Panel A: Trading frictions cluster					
IC1	*Trading frictions*	0.12	5.54	0.08	58.87
DC1	*Trading frictions (measured by volume)*	0.03	2.59	0.04	43.49
DC2	*Illiquidity*	0.06	2.58	0.08	42.66
DC3	*Trading frictions (measured by return)*	0.11	3.92	0.09	58.91
DC4	*Beta*	0.38	4.63	0.28	51.07
Panel B: Momentum cluster					
IC2	*Momentum*	0.65	4.23	0.53	44.47
DC5	*Short − run momentum*	0.13	3.11	0.15	46.00
DC6	*Long − run momentum*	0.71	3.49	0.70	38.46
Panel C: Other clusters					
IC3	*Value*	0.02	2.66	0.03	62.36
IC4	*Profitability*	0.31	2.61	0.41	31.01
IC5	*Intangibles*	0.31	2.20	0.49	27.51
IC6	*Investment*	0.13	1.96	0.24	41.93
DC7	*Profitability*	0.34	2.27	0.52	30.21
DC8	*Growth*	0.33	1.87	0.61	22.51
DC9	*Accruals*	0.11	1.90	0.20	28.22

Models	Performance of OS maximal Sharpe ratio portfolio			
	Mean(%)	S.D.(%)	Sharpe	MDD(%)
Panel A. S-WLS				
DC	0.56	1.36	1.42	8.37
IC	0.55	1.60	1.18	15.38
FF3	0.57	3.00	0.66	34.25
Q4	0.79	3.04	0.90	24.77
Car4	0.50	1.82	0.96	19.63
FF5	0.55	1.83	1.03	26.36
Lasso	0.87	8.09	0.37	50.64
Ridge	0.78	8.25	0.33	54.69
Enet	0.86	7.99	0.37	49.50

- Mean-variance efficiency (above right)
- Bayesian comparison (below left)

	$\kappa=2.0$	2.2	2.4	2.6	2.8	3.0	3.2	3.4	3.6	3.8	4.0
Panel A: Compare the DC-Model with each of benchmarks											
IC	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
FF3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Car4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Q4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
FF5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Panel B: Compare the DC-Model with benchmarks simultaneously											
DC	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
IC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FF3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Car4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FF5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Results: Performance of the C-IPCA model

IPCA (Kelly et al., 2019) performs well in the stock market.

- Assumption of IPCA: one risk exposure is a linear function of firm characteristics.

$$\beta_{k,t-1} = X_{t-1}\Gamma_k + u_{t-1}.$$

  - » $\beta_{k,t-1} \in R^N$: risk exposures of N stocks on $k\hat{}th$ latent factor.
  - » $\Gamma_k \in R^I$: loadings of all firm characteristics on $k\hat{}th$ risk exposure.
  - » $X_{t-1} \in R^{N*I}$: all firm characteristics of N stocks in month t-1.
- Adv. of IPCA.
  - Uses firm characteristics in the traditional PCA method.
- Disad. of IPCA.
  - Hard to be interpreted economically.
  - Many parameters to be estimated.

Improves of IPCA (C-IPCA).

- Assumptions: one risk exposure is a linear function of one cluster of firm characteristics.

$$\beta_{k,t-1} = X_{t-1}\Gamma_k + u_{t-1}.$$

  - $\Gamma_{k,i} = 0, if\ i\ is\ not\ in\ cluster\ k.$
- Similar to IPCA, uses firm characteristics in the traditional PCA method.
- Different from IPCA.
  - Easier to be interpreted economically.
  - Less parameters to be estimated.

Peformance of IPCA and C-IPCA

- 5 most important factors in C-IPCA.
  - Trading illiquidity, long-run momentum, short-run reversal, investment and market factor.
- Ability to explain each other's factors.

	Unadjusted		NW-Adjusted	
	C-IPCA	IPCA	C-IPCA	IPCA
Panel A: Alpha (t-statistic)				
C-IPCA1		0.56*** (3.88)		0.56*** (3.80)
C-IPCA2		0.53*** (2.86)		0.53** (2.40)
C-IPCA3		0.15 (1.05)		0.15 (0.91)
C-IPCA4		0.1 (0.73)		0.1 (0.54)
C-IPCA5		0.03 (1.65)		0.03 (1.56)
IPCA1	-0.07 (-1.66)		-0.07* (-1.82)	
IPCA2	0.01 (0.08)		0.01 (0.08)	
IPCA3	-0.02 (-0.2)		-0.02 (-0.15)	
IPCA4	0.07 (0.74)		0.07 (0.81)	
IPCA5	-0.04 (-0.32)		-0.04 (-0.32)	

Conclusions

We propose a new approach to construct factors from firm characteristics.

- Easier to be interpreted economically.
- Outperforms existing models.

We also apply our approach to the IPCA model of Kelly, Pruitt and Su (2019).

- Our factors are well linked to the associated economic risks.
- Can price assets no worse than the standard IPCA.

References

Cochrane, J. H. 2011. Presidential address: Discount rates. The Journal of finance 66:1047-1108.

Connor, G., & Korajczyk, R. A. (1988). Risk and return in an equilibrium APT: Application of a new test methodology. Journal of financial economics, 21(2), 255-289.

Feng, G., Giglio, S., & Xiu, D. (2020). Taming the factor zoo: A test of new factors. The Journal of Finance, 75(3), 1327-1370.

Freyberger, J., Neuhierl, A., & Weber, M. (2020). Dissecting characteristics nonparametrically. The Review of Financial Studies, 33(5), 2326-2377.

Han, Y., He, A., Rapach, D. and Zhou, G. 2021. Firm characteristics and expected stock returns. Available at SSRN.

Han, Y., He, A., Rapach, D. and Zhou, G. 2023. Cross-Sectional Expected Returns: New Fama-MacBeth Regressions in the Era of Machine Learning. Available at SSRN: https://ssrn.com/abstract=3185335 or http://dx.doi.org/10.2139/ssrn.3185335.

Harvey, C. R., Liu, Y., and Zhu, H. 2016. ... and the cross-section of expected returns. Review of Financial Studies 29:5–68.

Hou, K., Xue, C., and Zhang, L. 2015. Digesting anomalies: An investment approach. Review of Financial Studies 28:650-705.

Hou, K., Xue, C., and Zhang, L. 2015. Digesting anomalies: An investment approach. Review of Financial Studies 28:650-705.

Hou, K., Xue, C., & Zhang, L. (2020). Replicating anomalies. The Review of financial studies, 33(5), 2019-2133.

Kelly B, Pruitt S, and Su Y. 2019. Characteristics are covariances: a unified model of risk and return. Journal of Financial Economics 134:501–24.

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

McLean, R. D., & Pontiff, J. (2016). Does academic research destroy stock return predictability?. The Journal of Finance, 71(1), 5-32.

Stambaugh, R., and Yuan, Y. 2017. Mispricing factors. The Review of Financial Studies 30:1270-1315.

Acknowledgements

We are grateful to seminar participants at Washington University in St. Louis, conference participants at the 2022 Pacific Basin Finance, Economics, Accounting, and Management (PBFEAM), the 2023 Asia Meeting of the Econometric Society (AMES), the 2023 China Fintech Research Conference (CFTRC), the professors and phd students at the 2022 SoFiE Financial Econometrics Summer School, for their helpful comments.