Chimeras in the Factor Zoo

Zied Ben Chaouch, Andrew W. Lo, Manish Singh, and Katherine Xiong

MIT Laboratory for Financial Engineering (LFE), Computer Science and Artificial Intelligence Laboratory (CSAIL)

Objectives

- More than 150 anomalies have been reported within the cross-section of asset prices.
- We use machine learning (ML) methods to combine 150 factors into parsimonious models.
- In addition, we test the performance of our ML factor models on well-known anomaly portfolios and mutual fund returns.
- Finally, we compare the performance of our ML factors to the classical Fama-French and Hou-Xue-Zhang factor models.

Introduction

Factor models are routinely used in the financial industry to identify and quantify sources of systematic risk in order to manage the risk of a portfolio of securities or hedge investment positions, or in valuation contexts to estimate the cost of capital of an asset.

The *CAPM* [1, 2] simply identifies the systematic risk of a stock with its exposure to the market. However, the *CAPM* fails to explain the cross-section of asset prices and multiple anomalies (sources of systematic risk that are not captured by the *CAPM*) were documented [3]. More complex factor models were developed such as the Fama-French 3, 4, and 5-factors models (*FF3*, *FF5*, *FF6*) [4, 5, 6], the Carhart 4-factors model [7], or the 4 and 5 q-factors (*q4*, *q5*) by Hou, Xue, and Zhang [8, 9].

A large set of factors have been introduced in the literature to attempt to explain the cross-section of asset prices, however no single model has convincingly been able to capture most of the anomalies. Tools have been proposed to "tame the factor zoo" and compress the 150+ factors proposed in the literature into a parsimonious model. These techniques include the double selection *LASSO* approach [10] as well as *PCA* [11]. While the former consists of a model-selection approach, the latter aims to extract a set of "fundamental" latent factors. Our approach is similar to [11] as we aim to construct a set of latent factors that can explain away most of the anomalies in the cross section of asset prices.

Data

To construct the latent ML factors, we use the 150 well known factors replicated by [10] (publicly available). The data consists of monthly returns for each factor from July 1976 to December 2017. In summary, we have 150 factors, 498 time-series datapoints per factor, and no missing values. We normalize raw factor returns by their variance.

We test the performance of our factors on the following datasets (monthly returns over 1976-2017):

1. 75 Fama-French Portfolios from Kenneth French's online data library. We include the 25 Fama-French size and book-to-market portfolios, the 25 Fama-French size and operating profitability portfolios, and the 25 Fama-French size and investment portfolios.

2. 374 Hou-Xue-Zhang Portfolios. We use 187 anomalies across: momentum (41), value-versus-growth (32), investment (29), profitability (45), intangibles (30), and frictions (10). The $187 \times 2 = 374$ portfolios consist of the lowest and highest deciles for each anomaly.

3. 8,866 Mutual Fund Portfolios from the Refinitiv workspace (the Lipper databse) and from WRDS (the CRSP Mutual Funds database). Comprises of 8,866 mutual funds which cover 6 asset classes: Equity (4,480), Bond (2,242), Mixed Assets (1,132), Money Market (604), Alternatives (356), and Commodity (47).

Methods

We learn latent factor models using autoencoders (AEs) with different architectures. We instill economic intuition to the models by adding a mispricing penalty term to the MSE loss function:

$$\frac{1}{T \cdot N}\Sigma_{t,i}(r_{t,i} - \hat{r}_{t,i})^2 + (1+\gamma)\frac{1}{N}\Sigma_i(\bar{r}_i - \hat{\bar{r}}_i)^2. \quad (1)$$

After cross-validation, we set $\gamma = 10$ in all models.

ML models: simple one-layer AE, deep 2-layers AE, recursive AE, clustered AE (CAE, Figure 1), and recursive CAE. Recursive AEs reduce correlations among latent factors. Clustered AEs reduce over-representation in raw factors.

Methods (continued)

Recursive AEs: learn one latent factor at the time and remove the exposure to this factor.
CAE: cluster the 150 factors into 60 groups, connect inputs to their corresponding 60 nodes in the first hidden layer, and compress to latent factors.

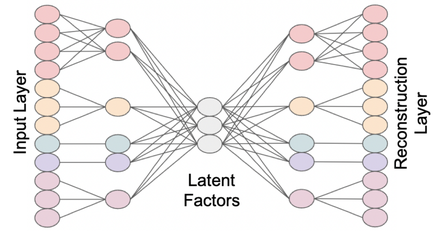

Figure 1:Architecture of the Clustered Autoencoder (CAE).

Activation Functions: linear (baseline) and tanh (nonlinear interactions).
Performance Evaluation: use linear, polynomial, and positive/negative time-series regressions. Standard errors are estimated through GMM.

- Time-series: R^2, not essential in asset pricing.
- Cross-section: "unexplained" excess return α (i.e., the intercept's t-stat is above 1.96).

Results

The latent factors perform better than the FF5 and q5 factor models on most test assets, however they do not always capture the market factor. Including a market factor improves their performance.

Model	Train	FF	HXZ	MF
FF5	63	80	58	69
q5	83	87	81	70
Linear 6	83 (81)	4 (48)	6 (82)	25 (68)
R-Linear 6	87 (85)	5 (72)	8 (84)	32 (70)
Tanh 7	86 (85)	52 (81)	68 (83)	69 (72)
R-Tanh 6	97 (94)	10 (61)	13 (82)	33 (70)
CAE 8	81 (82)	4 (53)	11 (77)	29 (67)
R-CAE 6	85 (85)	5 (75)	7 (85)	32 (70)
CAE Tanh 8	85 (83)	32 (60)	61 (77)	56 (68)

Table 1:Fraction of explained excess returns (in %). Brackets indicate the fraction obtained if we include a market factor.

Results (continued)

Latent factors can explain most FF5 factors except for the market factor. Models with non-linear activation functions are needed to explain q5 factors.

Model	FF5	q5
Linear 6	2*	3*
R-Linear 6	1*	3*
CAE 8	2*	3*
R-CAE 6	1*	3*
Tanh 7	0	1
R-Tanh 6	1*	3*
CAE Tanh 8	0	2

Table 2:Number of unexplained FF5 and q5 factors. (*) indicates that one of the unexplained factors is the market factor.

Conversely, FF5 factors fails to explain latent factors while q5 factors can only explain models with linear activation functions.

Model	FF5	q5	Lin	R-Lin	CAE	R-CAE
FF5	—	2	4	3	4	4
q5	0	—	1	2	1	3

Model	Tanh	R-Tanh	CAE Tanh
FF5	6	4	5
q5	4	2	3

Table 3:Number of factors unexplained by FF5 and q5.

References

[1] W. F. Sharpe, "Capital asset prices: A theory of market equilibrium under conditions of risk*," *The Journal of Finance*, vol. 19, no. 3, pp. 425–442, 1964.

[2] J. Lintner, "The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets," *The Review of Economics and Statistics*, vol. 47, no. 1, pp. 13–37, 1965.

[3] K. Hou, C. Xue, and L. Zhang, "Replicating Anomalies," *The Review of Financial Studies*, vol. 33, pp. 2019–2133, 12 2018.

[4] E. F. Fama and K. R. French, "Common risk factors in the returns on stocks and bonds," *Journal of Financial Economics*, vol. 33, no. 1, pp. 3–56, 1993.

[5] E. F. Fama and K. R. French, "A five-factor asset pricing model," *Journal of Financial Economics*, vol. 116, no. 1, pp. 1–22, 2015.

[6] E. F. Fama and K. R. French, "Choosing factors," *Journal of Financial Economics*, vol. 128, no. 2, pp. 234–252, 2018.

[7] M. M. Carhart, "On persistence in mutual fund performance," *The Journal of Finance*, vol. 52, no. 1, pp. 57–82, 1997.

[8] K. Hou, C. Xue, and L. Zhang, "Digesting anomalies: An investment approach," *The Review of Financial Studies*, vol. 28, pp. 650–705, 09 2014.

[9] K. Hou, H. Mo, C. Xue, and L. Zhang, "An Augmented q-Factor Model with Expected Growth*," *Review of Finance*, vol. 25, pp. 1–41, 02 2020.

[10] G. Feng, S. Giglio, and D. Xiu, "Taming the factor zoo: A test of new factors," *The Journal of Finance*, vol. 75, no. 3, pp. 1327–1370, 2020.

[11] M. Lettau and M. Pelger, "Estimating latent asset-pricing factors," *Journal of Econometrics*, vol. 218, no. 1, pp. 1–31, 2020.