Factor Strength and Factor Selection

An Application to U.S. Stock Market

Zhiyuan Jiang
28710967

Supervisors: Dr Natalia Bailey
Dr David Frazier

October 19, 2020

Motivation

Capital Asset Pricing Model (CAPM) is the benchmark of risk pricing. Assume we have:

- N asset: i = 1,2,3 $\cdots$ N
- K risk factors: j = 1,2,3 $\cdots$ K
- T observation: t = 1,2,3 $\cdots$ T

$$r_{it} - rf_t = a_i + \beta_{mt}(\bar{r}_t - rf_t) + \sum_{j=1}^{K} \beta_{ij} f_{jt} + \varepsilon_{it}$$

- r_{it}: asset's return
- r_{ft}: risk free return
- a_i: constant/intercept
- β_{mt}: market factor loading
- $\bar{r}_t$: market return
- β_{ij}: risk factor loading
- f_{jt}: risk factor
- ε_{it}: stochastic error

Why this matters

CAPM measures the risk and return relationship...

Fund managers will use this as reference when constructing portfolios.

- **Add factors to enhance risk pricing.**
- **New factors are booming**

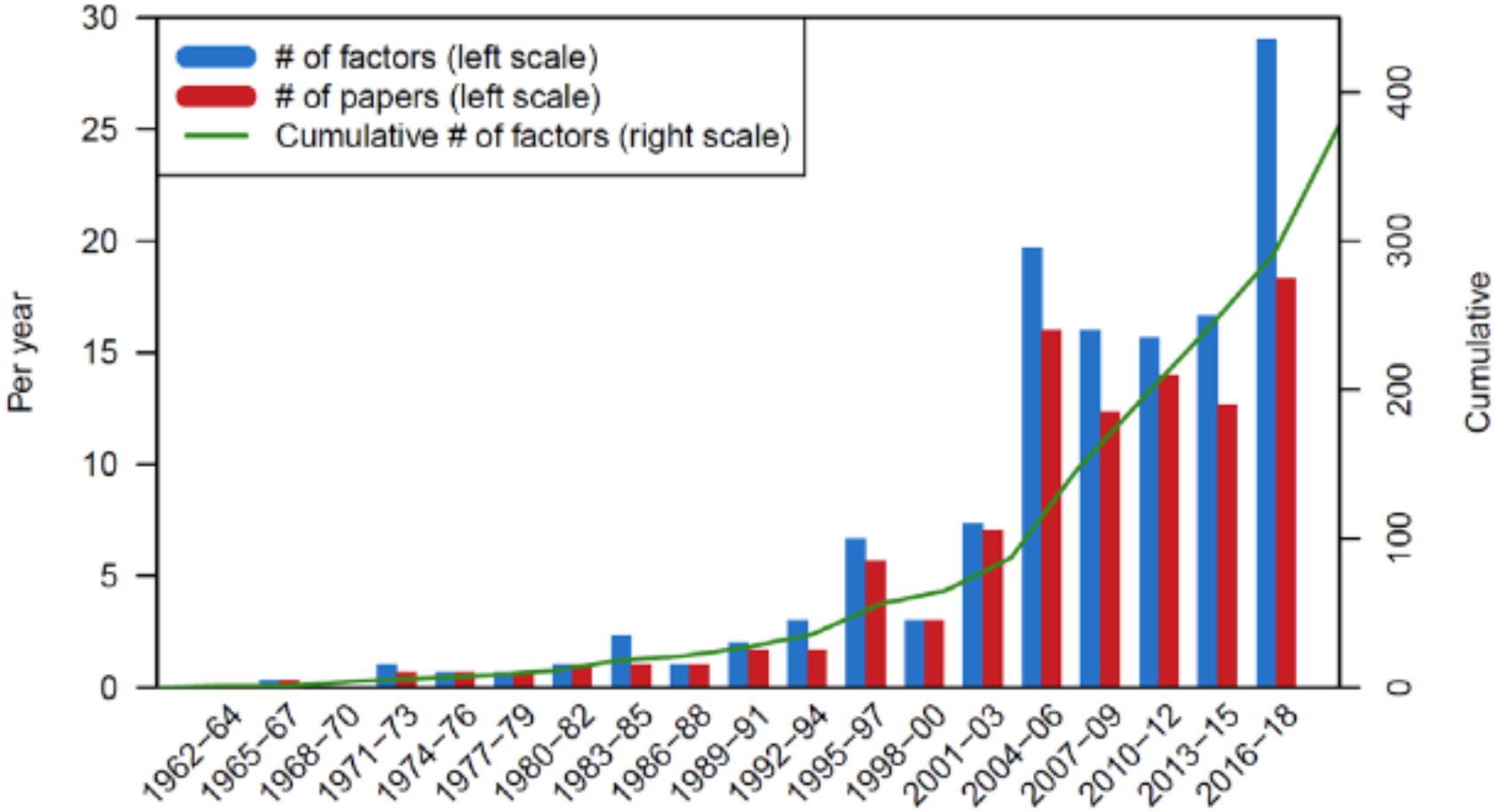


Figure: Factor amount growing through time.

(Harvey & Liu, 2019)

Problem inside factors

1. **Imposters**: The significant test is always under some confidence level. Some factors appear "significant" purely by chance.
   - Multiple testing problem
   - Can not replicate (?, ?)
2. **Results unreliable**: include "imposters" will distort the estimation
   - Statistical inference results unreliable (?, ?)
   - Inconsistent estimation (?, ?)
   - The portfolio may suffer loss because of this.

'We have a lot of questions to answer:
Firstly, which characteristics really provide ***independent*** *information about average returns? Which are subsumed by others ?'* Cochrane, 2011

Core Problem

How to select factors.

Core Problem

How to select factors.

Numerous methods have been proposed...

- Solving data mining problem
  - Bootstrapping (?, ?)
  - Bayes method(?, ?)
- Machine learning
  - Principle Component Analysis (PCA) (?, ?)
  - Lasso(?, ?)
- ...

Two Challenges

This project faces two challenges:

1. High dimensions of risk factor group.
   How to identify the significant risk factor. $\Rightarrow$ use factor strength as criterion.
2. Correlation among factors
   Traditional variable selection algorithm (Lasso) can not handle this.$\Rightarrow$ Will use elastic net techniques

Factor Strength

Stronger factor $\Rightarrow$ price more asset's risk $\Rightarrow$ generate more significant loadings β.
Factor strength defined in terms of number of non-zero significant loadings (Bailey, Kapetanios, & Pesaran, 2020).
Assume we have one factor with strength α_j:

$$|\beta_i| > 0, \quad i = 1, 2, 3, \cdots, [N^{\alpha_j}]$$
$$|\beta_i| = 0, \quad i = [N^{\alpha_j}] + 1, [N^{\alpha_j}] + 2, [N^{\alpha_j}] + 3, \cdots, N$$

Simply speaking: the more none-zero loadings a factor can generate, the stronger the factor is.

For every single risk factor, after running N regression against each assets, we will have a proportion:

$$\hat{\pi}_j = \frac{\text{Amount of significant non-zero loadings}}{\text{Amount of total loadings/N}}$$

$$\hat{\alpha}_j = \begin{cases} 1 + \frac{\ln \hat{\pi}_j}{\ln N}, & \text{if } \hat{\pi}_j > 0 \\ 0, & \text{if } \hat{\pi}_j = 0 \end{cases}$$

$$i = 1, 2, 3, \cdots N, \quad j = 1, 2, 3, \cdots K$$

$\alpha_j \in [0, 1]$.

0: no loadings are generate

1: the factor has significant non-zero loadings for each assets.

Elastic Net

Introduce by Zou and Hastie (2005), is a modified method to select factor.

For demonstration, we consider a simple linear regression.

$\boldsymbol{y} = c + \boldsymbol{x}\boldsymbol{\beta}$.

N observation pairs $(x_i,\ y_i\ )$ with $i = 1, 2, 3, \cdots N$, and $\boldsymbol{\beta} = [\beta_1, \beta_2, \beta_3, \cdots, \beta_K]'$.

$$\hat{\boldsymbol{\beta}} = \arg\min\{\sum_{i=1}^{N}(y_i - c - \sum_{j=1}^{K}\beta_j x_{ij})^2 + \phi P_\theta(\boldsymbol{\beta})\}$$

$$P_\theta(\boldsymbol{\beta}) = \sum_{j=1}^{K}[(1-\theta)\beta_j^2 + \theta|\beta_j|]$$

We use the R-package ***glmnet***. (Friedman, Hastie, & Tibshirani, 2010)

Parameter tuning

From the loss function, you can see we need to decided on two parameters: ϕ and θ.
I will not talk about the detail of the parameter tuning just for now...
The trick is: minimise the MSE (and cross-validation).

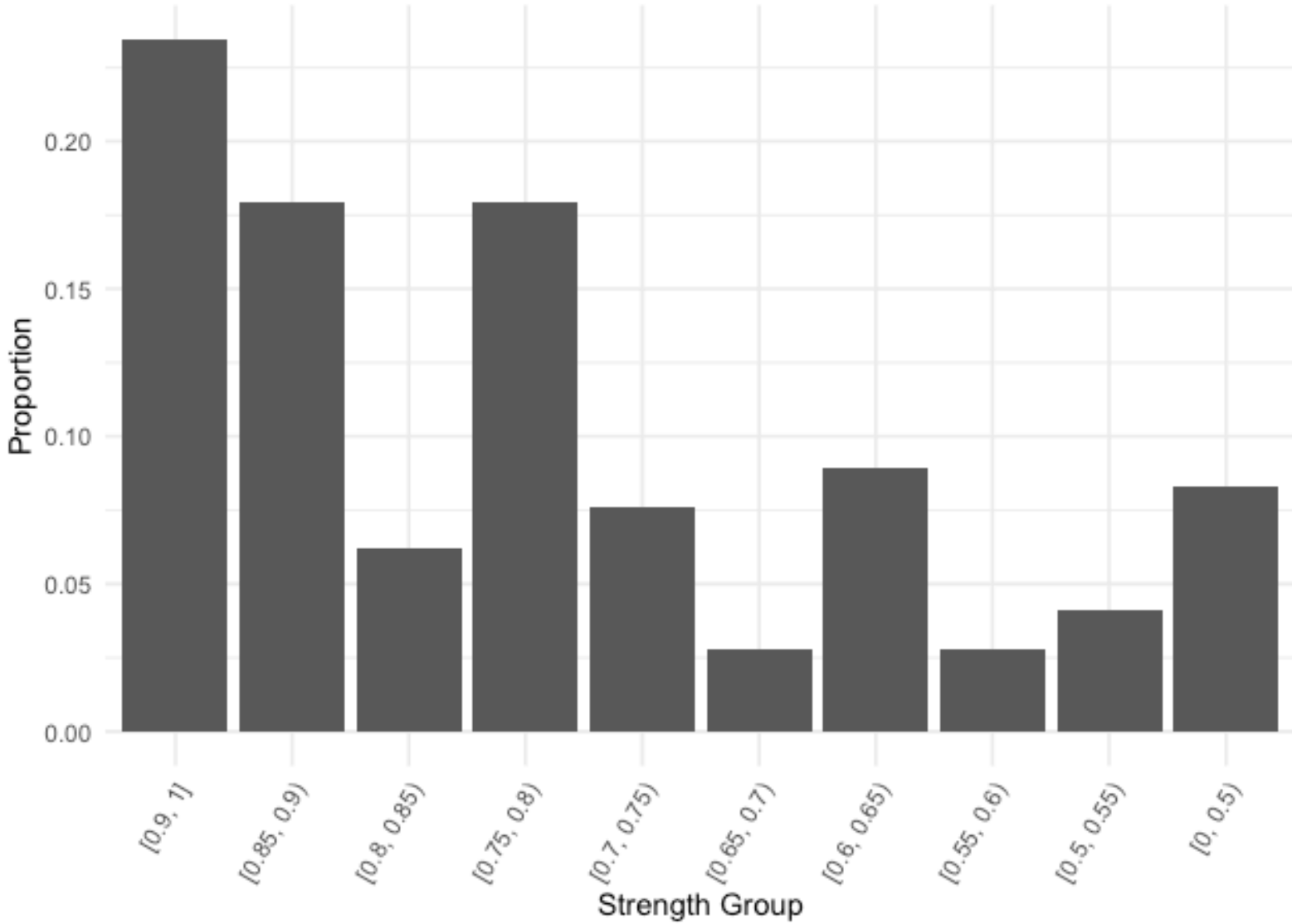
Data

The data set included two parts:

- **Assets**: Standard & Poor (S&P) 500 index companies, three year U.S. t-bill, and average market return.
- **Factor**: 145 factors plus the market factor
- **Time period**: Collect thirty years data: 1988:1-2017:12.
- We focusing on the thirty year data, but also use 10/20 year data set to see how the factor strength eveolove.

	Time Span	Number of Companies (n)	Observations Amount (T)
30 Years	January 1988 - December 2017	242	360

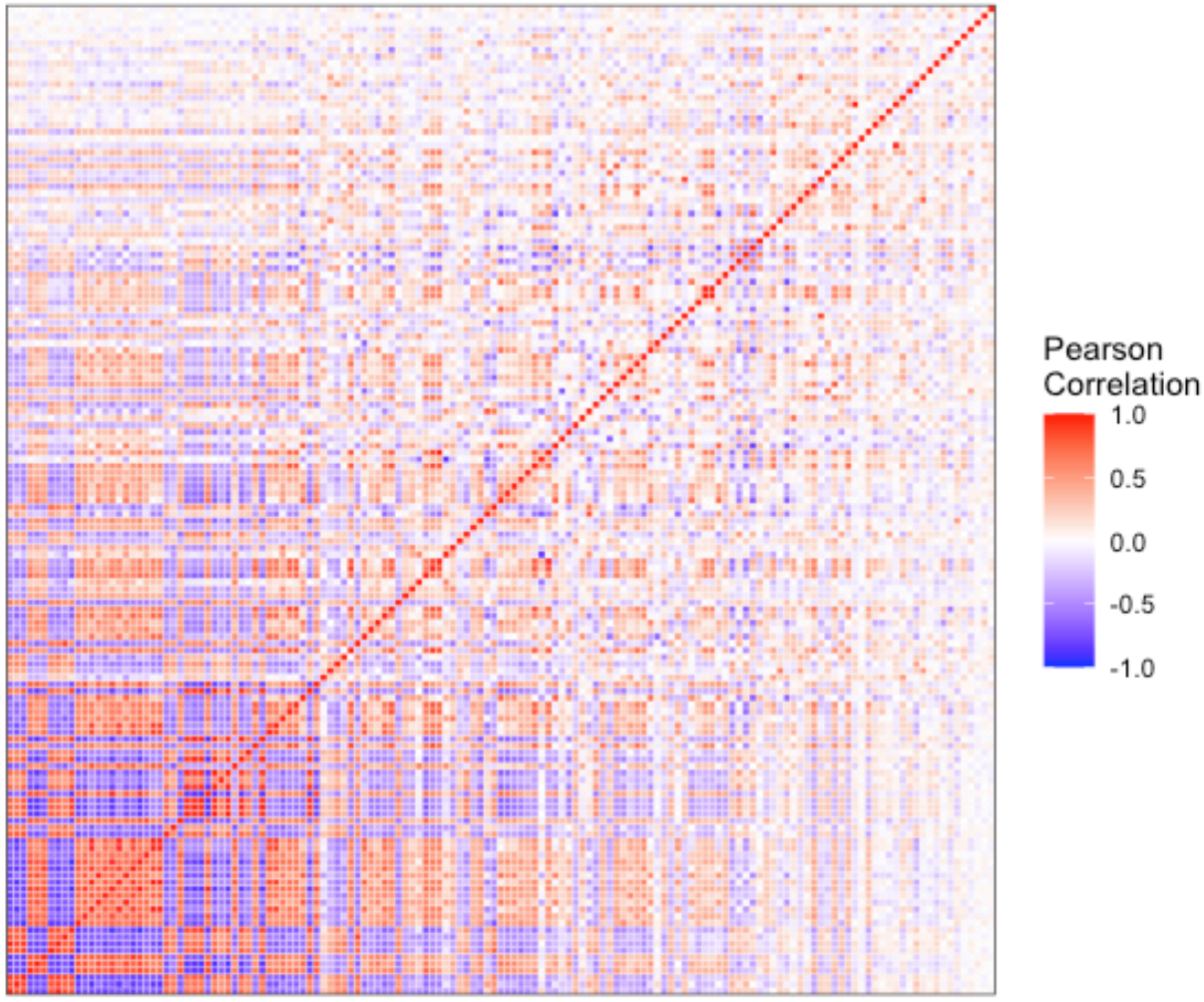
Proportion of factors with strength falling into different range (145 risk factors)



Top 10 strongest factors and three famous factors

	Thirty Year	
Rank	Factor	Strength
	Market	0.995
1	salecash	0.948
2	ndp	0.941
3	quick	0.940
4	age	0.940
5	roavol	0.938
6	ep	0.937
7	depr	0.935
8	cash	0.934
9	rds	0.931
10	dy	0.927
39	HML	0.894
68	SMB	0.804
96	UMD	0.745

Correlation of Factors: from strong to weak



Correlation among factors.

Factor Group	[0,0.5)	[0.5, 0.6)	[0.6, 0.7)	[0.7, 0.8)	[0.8,0.9)	[0.9,1]
Correlation Coefficient	0.0952	0.157	0.213	0.229	0.371	0.724
Factor Amount	12	10	17	37	35	34

- Correlation among strong factor is very high.
- Among weak factors is very low.
- Recall the problem Lasso can not handle correlation...

Factor Selection Result

Factor Group	[0,0.5)	[0.5, 0.6)	[0.6, 0.7)	[0.7, 0.8)	[0.8,0.9)	[0.9,1]	mix
Factor Amount	12	10	17	37	35	34	20
Proportion of Agreement (Exact)	68.7%	55.9%	42.8%	20.9%	17.7%	13.9%	34.6%
Proportion of Agreement (90%)	86.8%	72.0%	74.5%	72.0%	79.8%	74.4%	76.1%
Avg EN selection amount	2.11	4.47	8.67	14.67	13.51	12.37	8.45
Avg EN selection proportion	17.5%	44.73%	51.00%	39.65%	38.61%	36.38%	42.28%
Avg Lasso selection amount	2.06	3.87	8.43	13	12.19	10.46	7.26
Avg Lasso selection proportion	17.2%	38.76%	49.60%	35.14%	34.83%	30.75%	36.27%

- Agreement decrease with factor strength increase
- Lasso produce parsimonious model
- When facing weak factors, both Lasso and EN can well reduce redundancy.
- Eight of Top 10 most selected factors from mix factor group are strong factors.

Potential Extension

1. Using other criterion for tuning parameter
2. Categorised the factors and stocks
3. Using other methods to select factors, compare with the Lasso and Elastic net.

Thanks for listening

EN parameter tuning

Recall our simple toy multi-variables regression: $\boldsymbol{y} = c + \boldsymbol{x}\boldsymbol{\beta}$.
N observation pairs $(x_i,\ y_i\ )$ with $i = 1, 2, 3, \cdots N$, and $\boldsymbol{\beta} = [\beta_1, \beta_2, \beta_3, \cdots, \beta_K]'$.

$$\hat{\boldsymbol{\beta}} = \arg\min\{\sum_{i=1}^{N}(y_i - c - \sum_{j=1}^{K}\beta_j x_{ij})^2 + \phi P_\theta(\boldsymbol{\beta})\}$$

$$P_\theta(\boldsymbol{\beta}) = \sum_{j=1}^{K}[(1-\theta)\beta_j^2 + \theta|\beta_j|]$$

The R package ***glmnet*** provides function to tuning parameter ϕ, using cross-validation, targeting at minimise the MSE.
We use the same principle: minimise the MSE to determine our θ value.

Assume we have N units of stock, J risk factors, and T observations.

1. Assign first 90% of data as learning set, and rest 10% as test set.
2. Prepare a sequence of θ values, from 0 to 1, with step 0.01
3. For each θ, we use the learning set to fit a model, with ϕ selected by the function
4. Base on the fitted model, makes prediction and compare with the test set, and calculate the MSE.
5. The $\theta - \phi$ combination with smallest MSE is the winner.

In practice, because the problem of computation burden, we will randomly select 10 factors from each group, and 10 companies to conduct the procedure.
Then, we repeat the whole procedure 2000 times, and take the average of parameter results.

Bibliography I

Bailey, N., Kapetanios, G., & Pesaran, M. H. (2020). Measurement of factor strength: Theory and practice. *CESifo Working Paper*.

Cochrane, J. H. (2011, 8). Presidential address: Discount rates. *The Journal of Finance*, *66*, 1047-1108. Retrieved from http://doi.wiley.com/10.1111/j.1540-6261.2011.01671.x doi: 10.1111/j.1540-6261.2011.01671.x

Friedman, J., Hastie, T., & Tibshirani, R. (2010, 2). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*, 1-22. Retrieved from https://www.jstatsoft.org/index.php/jss/article/view/v033i01/v33i01.pdfhttps://www.jstatsoft.org/

Bibliography II

index.php/jss/article/view/v033i01 doi: 10.18637/jss.v033.i01

Harvey, C. R., & Liu, Y. (2019, 3). A census of the factor zoo. *SSRN Electronic Journal*. doi: 10.2139/ssrn.3341728

Zou, H., & Hastie, T. (2005, 4). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*, 301-320. Retrieved from http://doi.wiley.com/10.1111/j.1467-9868.2005.00503.x doi: 10.1111/j.1467-9868.2005.00503.x