Factor Strength and Factor Selection

An Application to U.S. Stock Market

Zhiyuan Jiang
28710967

Supervisors: Dr Natalia Bailey
Dr David Frazier

October 12, 2020

Motivation

Capital Asset Pricing Model (CAPM) is the benchmark of risk pricing.

$$r_{it} - r_{ft} = a_i + \beta_{im}(r_{mt} - r_{ft}) + \sum_{j=1}^{k} \beta_{ij} f_{jt} + \varepsilon_{it}$$

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

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

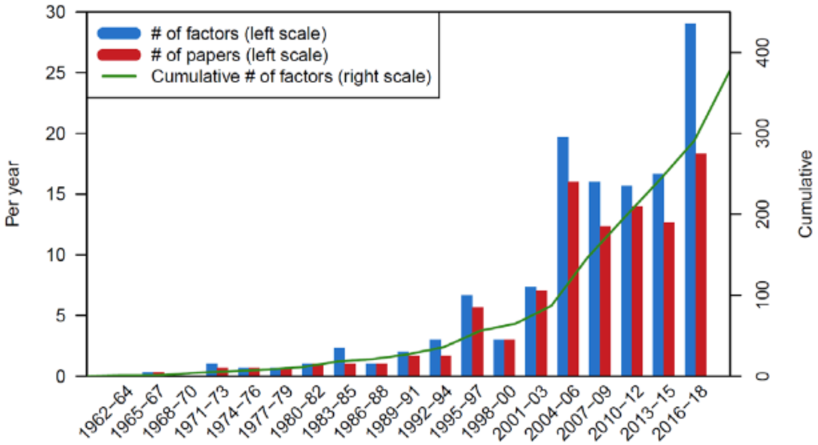


Figure: Factor amount growing through the year.

(Harvey & Liu, 2019)

Problem inside factors

1. **Imposters**: Some factors get positive results because of luck
   - multiple testing problem
   - Can not replicate (Hou, Xue, & Zhang, 2018)
2. **Results unreliable**: include "imposters" will distort the estimation
   - Statistical inference results unreliable (Gospodinov, Kan, & Robotti, 2017)
   - Inconsistent estimation (Anatolyev & Mikusheva, 2018)
   - ...

'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 research has been done...

- Solving data mining problem
- Bayes method
- Machine learning
- . . .

Two Challenges

This project faces two challenges:

1. High dimensions of data group
   How to identify the significant one? ⇒ use factor strength as criteria
2. Correlation among factors
   Traditional variable selection algorithm (Lasso) can not handle this.⇒ Will use elastic net techniques

Factor Strength

Strong factor $\Rightarrow$ price more asset's risk $\Rightarrow$ generate more significantly loadings β.

Factor strength is defined in terms of factor loading (Bailey, Kapetanios, & Pesaran, 2020).

Assume we have N different assets.

$$
\begin{aligned}
|\beta_j| &> 0, \quad j = 1, 2, 3, \cdots, [N^{\alpha_j}] \\
|\beta_j| &= 0, \quad j = [N^{\alpha_j}] + 1, [N^{\alpha_j}] + 2, [N^{\alpha_j}] + 3, \cdots, N
\end{aligned}
$$

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

For every single risk factor, after running a bunch of regression against different assets, we will have a proportion: Proportion $\hat{\pi}_n$, represent how many non-zero significant loadings are generated.

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

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

0 means no loadings are generate, and 1 means the factor can generate loadings to every assets.

Elastic Net

Introduce by Zou and Hastie (2005), is a improved method to select factor.
Considering the following loss function:

$$\hat{\beta}_{ij} = \arg\min_{\beta_{ij}}\{\sum_{i=1}^{n}[(r_{it} - r_{ft}) - \beta_{ij}f_{jt}]^2 + \lambda_2\sum_{i=1}^{n}\beta_{ij}^2 + \lambda_1\sum_{i=1}^{n}|\beta_{ij}|\}$$

The L_1 norm $\sum_{i=1}^{n}|\beta_{ij}|$ helps select the factor, reduce redundancy.
The L_2 norm $\sum_{i=1}^{n}\beta_{ij}^2$ helps handle the correlation.

Elastic Net: In empirical

We use R package ***glmnet***, and the package using loss function (Friedman, Hastie, & Tibshirani, 2010):

$$\hat{\beta}_i = \arg\min\{\frac{1}{2N}(x_{it} - \hat{a}_{iT} - \hat{\boldsymbol{\beta}}_{\boldsymbol{i}}'\boldsymbol{f_t}^2) + \phi P_\theta(\boldsymbol{\beta_i})\}$$

$$P_\theta(\boldsymbol{\beta_i}) = \sum_{j=1}^{k}[(1-\theta)\boldsymbol{\beta}_{ij}^2 + \theta|\boldsymbol{\beta}_{ij}|]$$

We have to decide two parameter: ϕ, and θ.

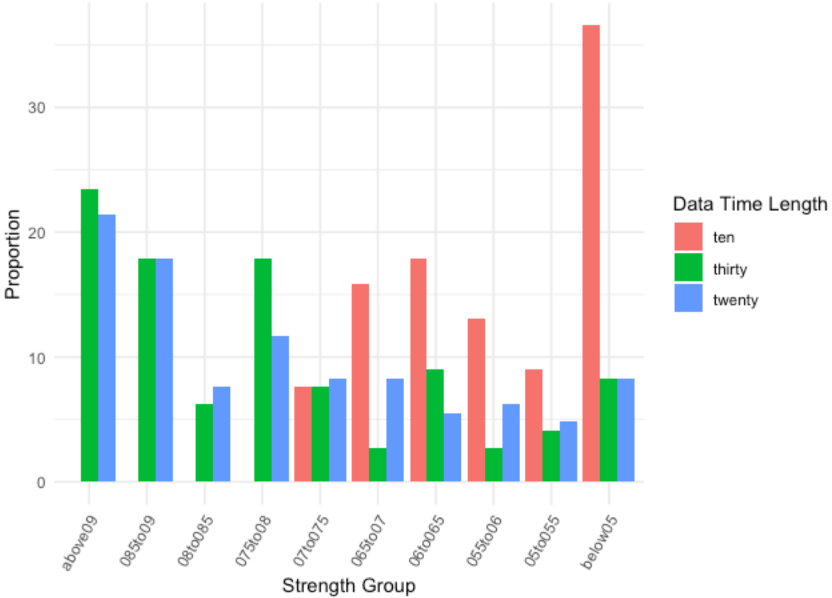
Data

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 one market factor
- **Time period**: Collect thirty years data: 1988:1-2017:12.
- Divided into three subsamples: 10/20/30 years.

	Time Span	Number of Companies (n)	Observations Amount (T)
10 Years	January 2008 - December 2017	419	120
20 Years	January 1998 - December 2017	342	240
30 Years	January 1988 - December 2017	242	360

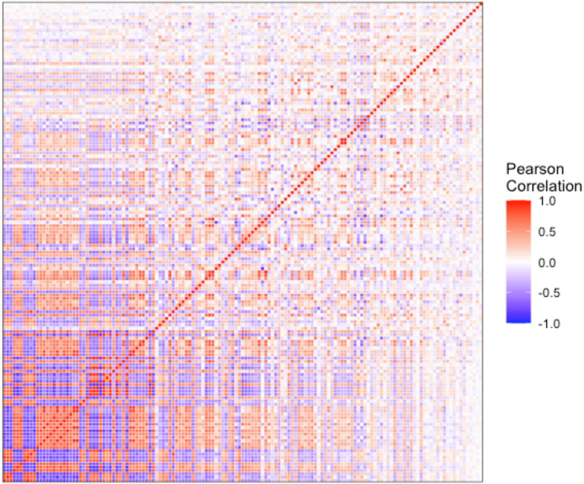
Proportion of Strength (145 risk factors)



Top 10 strong factors and three famous factors

Ten Year			Twenty Yera			Thirty Year		
Rank	Factor	Strength	Rank	Factor	Strength	Rank	Factor	Strength
	Market	0.988		Market	0.990		Market	0.995
1	beta	0.749	1	ndp	0.937	1	salecash	0.948
2	baspread	0.730	2	quick	0.934	2	ndp	0.941
3	turn	0.728	3	salecash	0.933	3	quick	0.940
4	zerotrade	0.725	4	lev	0.931	4	age	0.940
5	idiovol	0.723	5	cash	0.931	5	roavol	0.938
6	retvol	0.721	6	dy	0.929	6	ep	0.937
7	std_turn	0.719	7	roavol	0.929	7	depr	0.935
8	HML_Devil	0.719	8	zs	0.927	8	cash	0.934
9	maret	0.715	9	age	0.927	9	rds	0.931
10	roavol	0.713	10	cp	0.926	10	dy	0.927
20	UMD	0.678	29	HML	0.905	39	HML	0.894
24	HML	0.672	76	SMB	0.770	68	SMB	0.804
87	SMB	0.512	89	UMD	0.733	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

$$\hat{\boldsymbol{\beta}}_i = \arg\min\{\frac{1}{2N}(x_{it} - \hat{a}_{iT} - \hat{\boldsymbol{\beta}}_{\boldsymbol{i}}'\boldsymbol{f_t}^2) + \phi P_\theta(\boldsymbol{\beta_i})\}$$

$$P_\theta(\boldsymbol{\beta_i}) = \sum_{j=1}^{k}[(1-\theta)\boldsymbol{\beta}_{ij}^2 + \theta|\boldsymbol{\beta}_{ij}|]$$

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

Anatolyev, S., & Mikusheva, A. (2018, 7). Factor models with many assets: strong factors, weak factors, and the two-pass procedure. *CESifo Working Paper Series*. Retrieved from http://arxiv.org/abs/1807.04094

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

Bibliography II

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

Gospodinov, N., Kan, R., & Robotti, C. (2017, 9). Spurious inference in reduced-rank asset-pricing models. *Econometrica*, *85*, 1613-1628. doi: 10.3982/ecta13750

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

Bibliography III

Hou, K., Xue, C., & Zhang, L. (2018, 12). Replicating anomalies. *The Review of Financial Studies*, *33*, 2019-2133. Retrieved from `https://doi.org/10.1093/rfs/hhy131` doi: 10.1093/rfs/hhy131

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