Factor Selection and Factor Strength

Base on the U.S. Stock Market data

Research Plan

Zhiyuan Jiang

I.D:28710967

Supervisor : Dr. Natalia Bailey

Dr. Davide Frazier

May 17, 2020

Contents

1 Introduction and Motivation

Capital Asset Pricing Model (CAPM) (Sharpe (1964), Lintner (1965), and Black (1972)) has greatly changed the way people measure the relationship of asset's risk and return. Because the original model only contains one market factors, scholars after them are always trying to add new factors to create a multi-factor extension CAPM. The most famous example would be the three-factor model introduced by Fama and French (1992), they added the size and value factors into the original model. Carhart (1997) add the momentum factor to the three-factor models and thereby introduce the four-factor model. Researchers nowadays are still trying to find a sparse factor model. For instance, Kelly, Pruitt, and Su (2019) using the IPCA method construct a six-factor model and claim that their model outperforms all other existing sparse factor models.

In the other hand, the factor is abundant. Harvey and Liu (2019) had collected over 500 factors published in the top financial and economic journal, and the growth of new factors speed up since 2008. In his 2011 presidential address, J. H. Cochrane coined the term "factor zoo" to describe factor model is facing: researchers and practitioners are having too many options to help them pricing the risk.

The abundance of factors does not provides a rosy prospect to factor model. McLean and Pontiff (2016) found that once a new factor is published, the probability of using this factor to successfully predict the stock result will drop significantly. In recent research, Hou, Xue, and Zhang (2018) fail to replicate over 60% result relates to the factor model, which had been published in the financial and economic journal. The reliability of those factors has been doubt.

What's more, Harvey, Liu, and Zhu (2015) argue that the current threshold for the test statistic is too low, and this helps some factor yield a significant result purely out of luck.

The existence of some seemly strong, but de factor weak or even useless factor jeopardized the result of many research. Kan and Zhang (1999) found that if the Fama-MacBeth two-stage regression(Fama & MacBeth, 1973) want to yield meaningful and correct result, the structure of the model must be correctly identified. But involving some factor, which has no correlation with the cross-section return, will cause miss-identification and thereby mislead the researcher to have the wrong conclusion. Kleibergen's 2009 paper also argues that for FM regression, when the factor's loading is small, or the size of cross-section assets is large, the result of regressions will be incorrect.

Similar results with regards to weak or useless factor were also mentioned by other scholars. (see Kleibergen and Zhan (2015), Gospodinov, Kan, and Robotti (2017), and Anatolyev and Mikusheva (2018)) Due to all those reasons, (J. H. Cochrane, 2011) post the question: ”Which characteristic really provide independent information about average return. Which are subsumed by others ?”

To answer this question, lot’s of scholars and research had applied various methods to pick the best factors from the factor zoo. Such as Harvey and Liu (2017) introduced a bootstrap method. Pukthuanthong, Roll, and Subrahmanyam (2019) developed a protocol, trying to capture the real priced rick facto. Some other scholars are trying to use machine learning methods to reduce the potential candidates of useful factors, one stream of them are using a shrinkage and subset selection method like Lasso (Tibshirani, 1996) and its deviation to find the suitable factors (See, Rapach, Strauss, and Zhou (2013), Feng, Giglio, and Xiu (2019), Freyberger, Neuhierl, and Weber (2020)).

But for factors, especially in the high-dimension, correlation is common. J. Cochrane (2005) points out that the correlation between factors will drag the ability of using risk premium to infer factors. ne main drawback of Lasso regression is that, when Lasso is facing a group of variables which are correlated with each other, Lasso will only select on among all and does not care about which one it picks up. Therefore, a new technique called Elastic Net is created (Zou & Hastie, 2005), trying to deal with the selection of correlated variables.

For this project, we are trying to deal with the classical problem: what factors can have significant contributions to explain the asset risk and return relationship under the CAPM framework. And we go a step forward, take the correlation among all factors into the account. But before applying the Elastic Net method, we want to first investigate all those factor’s strength. Our interest is focused on first determine which factors have enough power to help us solve the risk pricing problem. Then, from this pre-determined relatively small factor group, we applied the Elastic Net method, trying to find out the most appropriate factors to help us form the multi-factor CAPM model.

2 Related Literature

This project is built on a series of literature about trying to use different factors to price risk. The Capital Asset Pricing Model developed by Sharpe (1964), Lintner (1965), and Black (1972) completely change the way people measure the relationship between risk and return. For the initial

CAPM models, only market factor been included, Fama and French (1992) develop the model into three-factors, Carhart (1997) added the momentum factors and created the four-factor models. Base on their three-factor model, Fama and French (2015) extend the model to five-factors, and recent research created a six-factors model (Kelly et al., 2019).

Kan and Zhang (1999) first illustrate the negative influence of including weak factor when applying FM two-stage regression (Fama & MacBeth, 1973). Kleibergen (2009), and Kleibergen and Zhan (2015) point out how a weak factor inside the multi-factor CAPM models would disguise the fact that some structure does not exist. Gospodinov et al. (2017) show how the involving of a spurious factor will distort statistical inference of parameters. Anatolyev and Mikusheva (2018) studied the behaviours of the model with the presence of weak factors under asymptotic settings. These findings provide us with the motivation to investigates the method of selecting strong factors.

This project also relates to some researches effort to identify useful factors from the factor zoo. Harvey et al. (2015) exam over 300 factors published on journals, presents that the traditional threshold for a significant test is too low for newly proposed factor, and they suggest to adjust the p-value threshold to around 3. McLean and Pontiff (2016) exams 97 different factors, find their ability of predict out of sample return declined, and thereby cast doubt on the reliability of some factors' pricing ability. Some other method like bootstrap (Harvey & Liu, 2017) or Bayesian procedure (Barillas & Shanken, 2018) were tried to find out useful factors from a large group of homogenous competitors. Pukthuanthong et al. (2019) discovered this problem from another prospect, they developed a protocol to identify whether a factor is indeed a priced factor or not.

This project also benefits from and will contribute to emerging literature concerning applying machine learning on selecting characteristics on predicting cross-section stock returns. Gu, Kelly, and Xiu (2020) elaborate on the benefit and advantages of using emerging machine learning algorithms in asset pricing. The advantages including more accurate predict result, and better efficiency. Various machine learning algorithms have been adopted on selecting factors for the factor model, especially in recent years. Lettau and Pelger (2020) applying Principle Components Analysis on investigating the latent factor of model. Lasso method, since it's ability to select features, is popular in the field of the factor selection. Rapach et al. (2013) applying the lasso regression, trying to find some characteristics from a large group to predict the global stock market's return. Feng et al. (2019) used the double-selected Lasso method (Belloni, Chernozhukov, & Hansen, 2014), a grouped lasso method (Huang, Horowitz, & Wei, 2010) is used by Freyberger et al. (2020) on picking factors from a group of candidates. Kozak, Nagel, and Santosh (2020) used a Bayesian-based method, combing with both Ridge and Lasso regression, argues that the factor sparse model is ultimately futile. The correlation underneath the factors, however, inevitably affect the choice of model. Due to the problem that Lasso model can not select from correlated factors, this project will apply the method called elastic net (Zou & Hastie, 2005) to select factors from a group of pre-choice factor candidates.

3 Methodology

3.1 Factor Strength

Capital Asset Pricing Model (CAPM), especially the model's multi-factor extension, has become the benchmark when studying the relationship between risk and return. Consider the following multi-factor models with stochastic error term ε_{it}:

$$r_{it} = c + \beta_0 x_m + \beta_{\mathbf{j}}{}' \mathbf{x_j} + \varepsilon_{it} \tag{1}$$

In the left-hand side, we have r_{it} denotes the return of security i at time t. In the other side, two vectors $\beta_{\mathbf{j}} = \{\beta_1, \beta_2, \cdots, \beta_J\}$ and $\mathbf{x_j} = \{x_1, x_2, \cdots, x_J\}$, represents J different factor loadings and J different factors respectively. x_m is the market factor, which usually represents by the difference between market expected return and the risk -free return. β_0 denotes the market factor loading. In general, CAPM and it's multi-factor extension divided an asset risk into two parts, the systematic part which can be captured by the market factor x_m, and the asset-specific idiosyncratic part which is demonstrated by different characteristics from the factor vector $\mathbf{x_j}$.

In order to reduce the dimension of the factors when using the Elastic Net, we use factor strength as a criterion to select some factors in ahead. The factor strength of factor x_j as α_j from Pesaran and Smith (2019), and Bailey, Kapetanios, and Pesaran (2020) is defined as the pervasiveness of a factor. The stronger a factor is, the more loading that factor generates against different assets will be significantly different from zero. For instance, if a factor has strength α, and base on the Fama-MacBeth two-stage regressions' first-stage regression, use this factor against N different assets, we will have $[N^{\alpha}]$ factor loading statistically significantly different from zero, here $[\cdot]$ operator will take the integer from the result:

$$\beta_j \neq 0, j = 1,2,3,\cdots,[N^{\alpha}]$$

$$\beta_j = 0, j = [N^{\alpha}]+1,[N^{\alpha}]+2,[N^{\alpha}]+3,\cdots,N$$

3.2 Elastic Net

Elastic net is a factor selection model introduced by Zou and Hastie (2005). The primary feature of the elastic net is that it has two penalty terms, combined the advantages of both ridge regression and lasso regression. Consider the model (1) with only one random factor x_j Assume that we are using the OLS to estimates the factor loading β_j OLS will trying to find $\beta_j$which has the smallest residual sum of squares.

$$\hat{\beta}_j = argmin\{\sum_{i=1}^{N}(r_{it} - \beta_{\mathbf{j}}\mathbf{x_j})^2\}$$

The elastic net, however, implies two different penalty terms when estimating the loading β_j.

$$\hat{\beta}_j = argmin\{\sum_{i=1}^{N}(r_{it} - \beta_{\mathbf{j}}\mathbf{x_j})^2 + \lambda_1 {\beta_{\mathbf{j}}}^2 + \lambda_2|\beta_{\mathbf{j}}|\} \tag{2}$$

The main advantage of elastic net, comparing with other factor selection method such as lasso, in this project is that elastic net can handle the problem of factor's correlation. Unlike lasso, the elastic net can choice factor properly from a group of related candidates.

4 Preliminary Result

In current stage, we have only studied the property of factor strength α under finite sample scenario. In purpose of this, we have designed and applied a Monte Carlo Simulation. The design details and result table can be seen at the Appendix A and Appendix B

To measure the goodness of simulation, we calculate the difference between the estimated factor strength and assigned factor strength as bias. Base on the bias,, we also calculated the Mean Squared

Error (MSE) for each setting.

From the result, we can easily find out that the error converge to zero when the strength α increases. When the $\alpha_x = 1$, we obtain the unbiased $\hat{\alpha}_x$

In the other hand, when the α is at a relatively low level, the estimation result will tend to overestimate the strength, and the level of overestimation decrease with the actual strength increase.

5 Further Plan

For the next step of this project, we will start the empirical analyse.

We will use companies return from Standard & Poors (S&P) 500 index as assets, to exam factors from Harvey and Liu (2017)'s factor list. The time span will be 30 years.

For the purpose of evaluation the selecting factors, we are planing using Out of Sample (OOS) method to predict the future return, and therefore exam how good those selected factors are.

References

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*.

Barillas, F., & Shanken, J. (2018, 4). Comparing asset pricing models. *The Journal of Finance*, *73*, 715-754. Retrieved from http://doi.wiley.com/10.1111/jofi.12607 doi: 10.1111/jofi.12607

Belloni, A., Chernozhukov, V., & Hansen, C. (2014, 4). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, *81*, 608-650. doi: 10.1093/restud/rdt044

Black, F. (1972). Capital market equilibrium with restricted borrowing. *The Journal of Business*, *45*, 444-455. Retrieved from www.jstor.org/stable/2351499

Carhart, M. M. (1997, 3). On persistence in mutual fund performance. *The Journal of Finance*, *52*, 57-82. Retrieved from http://doi.wiley.com/10.1111/j.1540-6261.1997.tb03808.x doi: 10.1111/j.1540-6261.1997.tb03808.x

Cochrane, J. (2005). *Asset pricing*.

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

Fama, E. F., & French, K. R. (1992, 6). The cross-section of expected stock returns. *The Journal of Finance*, *47*, 427-465. Retrieved from http://doi.wiley.com/10.1111/j.1540-6261.1992.tb04398.x doi: 10.1111/j.1540-6261.1992.tb04398.x

Fama, E. F., & French, K. R. (2015, 4). A five-factor asset pricing model. *Journal of Financial Economics*, *116*, 1-22. doi: 10.1016/j.jfineco.2014.10.010

Fama, E. F., & MacBeth, J. D. (1973, 5). Risk, return, and equilibrium: Empirical tests. *Journal of*

Political Economy, *81*, 607-636. doi: 10.1086/260061

Feng, G., Giglio, S., & Xiu, D. (2019, 1). *Taming the factor zoo: A test of new factors.* Retrieved from `http://www.nber.org/papers/w25481.pdf` doi: 10.3386/w25481

Freyberger, J., Neuhierl, A., & Weber, M. (2020, 4). Dissecting characteristics nonparametrically. *The Review of Financial Studies*, *33*, 2326-2377. Retrieved from `https://doi.org/10.1093/rfs/hhz123` doi: 10.1093/rfs/hhz123

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

Gu, S., Kelly, B., & Xiu, D. (2020, 2). Empirical asset pricing via machine learning. *The Review of Financial Studies*, *33*, 2223-2273. Retrieved from `https://doi.org/10.1093/rfs/hhaa009` doi: 10.1093/rfs/hhaa009

Harvey, C. R., & Liu, Y. (2017, 12). False (and missed) discoveries in financial economics. *SSRN Electronic Journal*. doi: 10.2139/ssrn.3073799

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

Harvey, C. R., Liu, Y., & Zhu, H. (2015, 10). … and the cross-section of expected returns. *The Review of Financial Studies*, *29*, 5-68. Retrieved from `https://doi.org/10.1093/rfs/hhv059` doi: 10.1093/rfs/hhv059

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

Huang, J., Horowitz, J. L., & Wei, F. (2010, 8). Variable selection in nonparametric additive models. *Annals of Statistics*, *38*, 2282-2313. doi: 10.1214/09-AOS781

Kan, R., & Zhang, C. (1999, 2). Two-pass tests of asset pricing models with useless factors. *The Journal of Finance*, *54*, 203-235. Retrieved from `http://doi.wiley.com/10.1111/0022-1082.00102` doi: 10.1111/0022-1082.00102

Kelly, B. T., Pruitt, S., & Su, Y. (2019, 12). Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics*, *134*, 501-524. doi: 10.1016/j.jfineco.2019.05.001

Kleibergen, F. (2009, 4). Tests of risk premia in linear factor models. *Journal of Econometrics*, *149*, 149-173. doi: 10.1016/j.jeconom.2009.01.013

Kleibergen, F., & Zhan, Z. (2015, 11). Unexplained factors and their effects on second pass r-squared's. *Journal of Econometrics*, *189*, 101-116. doi: 10.1016/j.jeconom.2014.11.006

Kozak, S., Nagel, S., & Santosh, S. (2020, 2). Shrinking the cross-section. *Journal of Financial Economics*, *135*, 271-292. doi: 10.1016/j.jfineco.2019.06.008

Lettau, M., & Pelger, M. (2020, 2). Estimating latent asset-pricing factors. *Journal of Econometrics*. doi: 10.1016/j.jeconom.2019.08.012

Lintner, J. (1965). The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *The Review of Economics and Statistics*, *47*, 13-37. doi: 10.2307/1924119

McLean, R. D., & Pontiff, J. (2016, 2). Does academic research destroy stock return predictability? *The Journal of Finance*, *71*, 5-32. Retrieved from `http://doi.wiley.com/10.1111/jofi.12365` doi: 10.1111/jofi.12365

Pesaran, M. H., & Smith, R. P. (2019). The role of factor strength and pricing errors for estimation and inference in asset pricing models. *CESifo Working Paper Series*.

Pukthuanthong, K., Roll, R., & Subrahmanyam, A. (2019, 8). A protocol for factor identification. *Review of Financial Studies*, *32*, 1573-1607. Retrieved from `https://doi.org/10.1093/rfs/hhy093` doi: 10.1093/rfs/hhy093

Rapach, D. E., Strauss, J. K., & Zhou, G. (2013, 8). International stock return predictability: What is the role of the united states? *The Journal of Finance*, *68*, 1633-1662. Retrieved from `http://doi.wiley.com/10.1111/jofi.12041` doi: 10.1111/jofi.12041

Sharpe, W. F. (1964, 9). Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance*, *19*, 425-442. Retrieved from `http://doi.wiley.com/10.1111/j.1540-6261.1964.tb02865.x` doi: 10.1111/j.1540-6261.1964.tb02865.x

Tibshirani, R. (1996, 1). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*, 267-288. Retrieved from `http://doi.wiley.com/10.1111/j.2517-6161.1996.tb02080.x` doi: 10.1111/j.2517-6161.1996.tb02080.x

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

A Monte Carlo Design

In this section, I will introduce the baseline design setting of the Monte Carlo Simulation and provides a preliminary result of the simulation.

A.1 Monte Carlo Design

Before start using the real data, we want to study the property of α by running Monte Carlo simulation and in this section, I will introduce the basic simulation design.

Consider the following model with stochastic error:

$$r_{it} = f_1(\bar{r}_t - r_f) + f_2(\theta_i x_t) + \varepsilon_{it} \tag{2}$$

In this Monte Carlo simulation, we consider a dataset has $i = 1, 2, \ldots, n$ different assets, with $t = 1, 2, \ldots, T$ different observations. $j = 1, 2, \ldots, k$ different factors and one market factors are included in the simulation.

$f_1(\cdot)$ and $f_2(\cdot)$ are two different functions represent the unknown mechanism of market factor and other factors in pricing asset risk. $(\bar{r}_t - r_f)$ is the market return, calculated from market or index return $\bar{r}_t$ minus risk free return r_f. r_{it} is the stock return, θ_{jt} denotes factors other than market factors and $\beta_{ij}$is the corresponding factor loading. ε_{it} is random error with structure can be defined in different designs. Notice that the β_{ij} will be influenced by each factor's strength α_j, where we have α as defined in section 3.1. And for each factor, we assume they follow a multinomial distribution with mean zero and a $k \times k$ variance-covariance matrix Σ. The diagonal of matrix Σ indicates the variance of each factor, and the rest represent the correlation among all k factors. In this model, we can control several parts to investigates different scenarios of the simulation:

A.2 Baseline Design

Follow the model (2), we assume both $f_1(a)$ and $f_2(a)$ are linear function:

$$f_1(a) = c_i + \beta a$$

$$f_2(a) = a$$

Therefore, the model with single factor can be write as:

$$r_{it} = c_i + \theta_i x_t + \varepsilon_{it}$$

The constantc_i is generated from a uniform distribution $U[-0.5, 0.5]$. θ_i is the factor loading, and x_t is factor with strength α_x. To generate factors loading, we employed a two steps strategy. First we generate a whole factor loadings vector $\theta_{\mathbf{i}} = (\theta_{i1}, \theta_{i2} \cdots, \theta_{in})$, All elements of the vector follows $IIDU(\mu_\theta - 0.2, \mu_\theta + 0.2)$. The μ_θ has been equalled to 0.71 to ensure all values apart from zero. After generating the vector, we randomly selected $[n^{\alpha_x}]$ elements from $\theta_{\mathbf{i}}$ to keep their value and set the other elements to zero. This step ensures the loading reflects the strength of each factor. For the stochastic error term, in this baseline design, we assume it follows a Standard Gaussian distribution, but we can easily extend it into a more complex form.

Follow the same idea, we also construct a two factor model:

$$r_{it} = c_i + \lambda x_m + \theta_i x_t + \varepsilon_{it}$$

Here the x_m is the market factor which assumably has strength $\alpha_m = 1$. λ is the market factor loading as a vector with all elements different from zero.

For each of the those different models, we consider the $T = \{120, 240, 360\}$, $n = \{100, 300, 500\}$. The market factor will have strength $\alpha_m = 1$ all the time, and the strength of the other factor in two factor model will be $\alpha_x = \{0.5, 0.7, 0.9, 1\}$. For every setting, we will replicate 500 times independently, all the constant c_i and loading θ_i will be re-generated for each replication.

B Simulation Result Table

Table 1: Simulation result of single factor model

	Single Factor					
	Biass			MSE		
$\alpha = 0.5$						
n \ T	120	240	360	120	240	360
100	0.194	0.188	0.199	0.050	0.047	0.053
300	0.224	0.224	0.226	0.062	0.062	0.062
500	0.229	0.237	0.225	0.064	0.067	0.062
$\alpha = 0.7$						
100	0.093	0.090	0.092	0.013	0.012	0.013
300	0.101	0.098	0.101	0.014	0.008	0.014
500	0.101	0.107	0.100	0.015	0.015	0.014
$\alpha = 0.9$						
100	0.023	0.022	0.023	0.001	0.001	0.001
300	0.023	0.023	0.024	0.001	0.001	0.001
500	0.023	0.023	0.024	0.001	0.001	0.001
$\alpha = 1.0$						
100	0.000	0.000	0.000	0.000	0.000	0.000
300	0.000	0.000	0.000	0.000	0.000	0.000
500	0.000	0.000	0.000	0.000	0.000	0.000

Table 2: Simulation result of two factor model

	Two Factor					
	Biass			MSE		
$\alpha_x = 0.5$, $\alpha_m = 1.0$						
n \ T	120	240	360	120	240	360
100	0.221	0.219	0.221	0.050	0.049	0.050
300	0.253	0.253	0.253	0.042	0.064	0.065
500	0.268	0.266	0.269	0.072	0.071	0.071
$\alpha_x = 0.7$, $\alpha_m = 1.0$						
100	0.100	0.101	0.100	0.010	0.010	0.010
300	0.113	0.113	0.112	0.013	0.013	0.013
500	0.118	0.118	0.119	0.014	0.014	0.014
$\alpha_x = 0.9$, $\alpha_m = 1.0$						
100	0.024	0.023	0.024	0.001	0.001	0.001
300	0.025	0.025	0.025	0.001	0.001	0.001
500	0.026	0.025	0.025	0.001	0.001	0.001
$\alpha_= 1.0$, $\alpha_m = 1.0$						
100	0.000	0.000	0.000	0.000	0.000	0.000
300	0.000	0.000	0.000	0.000	0.000	0.000
500	0.000	0.000	0.000	0.000	0.000	0.000