Factor Selection and Factor Strength

Base on the U.S. Stock Market data

Research Plan

Zhiyuan Jiang
I.D:28710967

May 12, 2020

Contents

1 Introduction and Motivation **3**

2 Methodology **3**

2.1 Factor Strength: Intuition 4

2.2 Monte Carlo Design 4

3 Preliminary Analyse Result **5**

3.1 Baseline Design 5

3.2 Simulation Result 6

References **7**

A Simulation Result Table **8**

1 Introduction and Motivation

Capital Asset Pricing Model (CAPM), created by Sharpe (1964) and Lintner (1965) is one of the most famous and wildly used model on explaining the relationship between financial asset's risk and return. Many scholars are trying to find and add new variables to the CAPM model to enhance its ability to capturing the dynamics between asset return and return volatility. One of the most famous examples of new factors are the size factor (SMB) and book-to-market factor(HML) found by Fama and French (1992)

A recent study by Harvey and Liu (2019), revealed the fact that in recent years new factors and paper illustrate how those factors are helping explain the relationship between risk and return were in abundance. In their 2015 papers, Harvey, Liu, and Zhu coined the term "factor zoo" which precisely described the fact that the field of financial economics has too many factors. And some, if not most, of them, can provide seemly significant results purely because of luck. Harvey (2017) provides some insight and explanation about this phenomena from the prospect of critical value and the journal's preference.

Inside the factor zoo, it is worth noticing that the ability of pricing risk is various among factors. Some researches (see Kleibergen, 2009, and Gospodinov, Kan, & Robotti, 2017) argues that for a weak factor which has small coefficient or even no coefficient, the statistic inference for a model containing such factors will be unreliable. Kan and Zhang (1999) warned that when including a factor has no correlation with the asset return in a two-pass method of testing the pricing model, the model will falsely identified that useless model as significant more usually than it should. limited explaining power before contains it into the CAPM becomes crucial. In their recent paper, Bailey, Kapetanios, and Pesaran (2020) introduce a new framework to measure a factor's ability or strength of pricing the risk of assets. We will briefly review their idea in the following section 2.1

In this project, I adopt the framework provided by Bailey et al. (2020) to exam the strength of factors from Harvey and Liu's factor zoo, base on the securities return of Standard & Poors (S&P) 500 index companies. And then, by applying some machine learning techniques and algorithms, I want to select strong factors which can most accurately pricing the asset's risk.

2 Methodology

In this section, I will briefly introduce the intuition behind Bailey et al.'s factor strength, and then provides a framework about how the idea of factor strength implied on a Monte Carlo simulation.

2.1 Factor Strength: Intuition

Generally speaking, the strength of factor represents how pervasiveness the factor is. The stronger a factor is, the more stock it can significantly influence at a point of time. Assume a single factor CAPM model:

$$y_{it} = \beta_i + \theta_i x_t + \varepsilon_{it} \tag{1}$$

In this model, we collect n different assets from T different time point. The left hand side y_{it} is the excess return for asset $i, \{i = 1, 2, \cdots n\}$ at time $t, \{t = 1, 2, \cdots, T\}$. θ_i here represents the factor loading of factor x_t at time t. After running the OLS regression between the asset's return and factor, we can collect a bunch of different loading θ_i, we can state that:

$$|\theta_i| > c_p(n) \quad i = 1, 2, \ldots, [n^\alpha]$$

$$|\theta_i| = 0 \quad i = [n^\alpha] + 1, [n^\alpha] + 2, \ldots, n$$

Here $c_p(n)$ denotes a multi-test corrected critical value for t-test. Those two equation clearly demonstrate how factor strength α is defined intuitively. For the first $[n^\alpha]$ term, where $[\cdot]$ will take integer part, the factor loading will statistically significant different from 0, and the rest loadings will not have any significant influence on pricing asset risk. For a factor has strength $\alpha = 1$, that factor will be significant for every assets at the same time. The more observation the factor can significantly influence, the stronger the factor is, and vice versa.

2.2 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 2.1. 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:

3 Preliminary Analyse Result

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

3.1 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.

3.2 Simulation Result

The detailed simulation result table has been showing in the Appendix A

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 $\hat{\alpha}_x$

References

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

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

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. (2017, 1). The scientific outlook in financial economics. *SSRN Electronic Journal*. doi: 10.2139/ssrn.2893930

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

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

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

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

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

A Simulation Result Table

Table 1: Simulation result of single factor model

	Single Factor					
	Biass			MSE		
$\alpha = 0.5$						
T n	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		
$\alpha_x = 0.5$, $\alpha_m = 1.0$						
T n	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