1 Factor Strength

The concept of factor strength employed by this project comes from Bailey, Kapetanios, and Pesaran (2020), and it was first introduced by Bailey, Kapetanios, and Pesaran (2016). They defined the strength of factor from prospect of the cross-section dependences and connect it to the pervasiveness of the factor, which is captured by the factor loadings. The method present in the initial paper is limited, since the factor strength can only be consistently estimated when the strength is higher than 0.5. In a latter paper, Bailey, Kapetanios, and Pesaran (2019) extended the method present before bt loosen some restrictions, and further developed into the method we employed in this project (Bailey et al., 2020).

1.1 Definition

Consider the following multi-factor model for n different cross-section units and T observations with k factors.

$$x_{it} = a_t + \sum_{j=1}^{k} \beta_{ij} f_{jt} + \varepsilon_{it} \tag{1}$$

In the left-hand side, we have x_{it} denotes the cross-section unit i at time t, where $i = 1, 2, 3, \cdots, n$ and $t = 1, 2, 3, \cdots T$. In the other hand, a_i is the constant term, which does not variate through the time. f_{jt} of $j = 1, 2, 3 \cdots k$ is factors included in the model, and β_{ij} is the corresponding factor loading. ε_{it} is the stochastic error term.

The factor strength is dependent on how many non-zero loadings a factor can generate. For factor f_{jt} with n different factor loading β_{ij}, we assume that:

$$|\beta_j| > 0 \quad i = 1, 2, \ldots, [n^{\alpha_j}]$$

$$|\beta_j| = 0 \quad i = [n^{\alpha_j}] + 1, [n^{\alpha_j}] + 2, \ldots, n$$

The α_j represents strength of facto f_{jt} and $\alpha_j \in [0, 1]$. If factor has strength α_j, we will assume that the first $[n^{\alpha_j}]$ loadings are all different from zero, and here $[\cdot]$ is defined as integral operator, which will only take the integral part of inside value.The rest $n - [n^{\alpha_j}]$ terms are all equal to zero. Assume for a factor which has strength $\alpha = 1$, the factor's loadings will be non-zero for all cross-section units. We will refer such factor as strong factor. And if we have factor strength $\alpha = 0$, it means that the factor cannot generate any loading different from zero, and we will describe such factor as useless factor of. For any factor with strength in [0.5, 1], we will refer them as semi-strong factor. In general term, the more non-zero loading a factor can generate, the stronger the factor's strength is.

1.2 Estimation

To estimate the strength α_j, Bailey et al. (2020) provides following estimation.

Here we consider a simplified model (1), a factor model with only one factor named f with different value f_t at different time. β_{it} is the factor loading of unit i at time t. v_{it} is the stochastic error term.

$$x_{it} = a_i + \beta_{it} f_t + v_{it} \tag{2}$$

Assume we have n different assets and T observations for each assets: $i = 1, 2, 3, \cdots, n$ and $t = 1, 2, 3, \cdots T$. Running the OLS regression for each $i = 1, 2, 3 \cdots, n$, we obtain:

$$x_{it} = \hat{a}_{iT} + \hat{\beta}_{iT} f_t + \hat{v}_{it}$$

For every factor loading $\hat{\beta}_{iT}$, we can exam their significance by constructing a t-test. The t-test statistic will be $t_{iT} = \frac{\hat{\beta}_{iT} - 0}{\hat{\sigma}_{iT}}$. Then the test statistic for the corresponding $\hat{\beta}_i$ will be:

$$t_{iT} = \frac{(\mathbf{f}'\mathbf{M}_\tau \mathbf{f})^{1/2} \hat{\beta}_{iT}}{\hat{\sigma}_{iT}} = \frac{(\mathbf{f}'\mathbf{M}_\tau \mathbf{f})^{-1/2} (\mathbf{f}'\mathbf{M}_\tau \mathbf{x}_i)}{\hat{\sigma}_{iT}} \tag{3}$$

Here, the $\mathbf{M}_\tau = \mathbf{I}_T - T^{-1} \tau \tau'$, and the τ is a $T \times 1$ vector with every elements equals to 1. $\mathbf{f}$ and $\mathbf{x_i}$ are two vectors with: $\mathbf{f} = (f_1, f_2 \cdots, f_T)'$ $\mathbf{x_i} = (x_{i1}, x_{i2}, \cdots, x_{iT})$. The denominator $\hat{\sigma}_{iT} = \frac{\sum_{i=1}^{T} \hat{v}_{it}^2}{T}$.

Using this test statistic, we then defined an indicator function as: $\ell_{i,nT} := \mathbf{1}[|t_{it}| > c(n)]$. If the t-statistic t_{iT} is greater than certain critical value $c_p(n)$, $\hat{\ell}_{i,nT} = 1$. In other word, we will count one if the factor loading $\hat{\beta}_{ij}$ is significant. With the indicator function, we then defined π_{nT} as the proportion of significant factor loading amount to the total factor loadings amount:

$$\hat{\pi}_{nT} = \frac{\sum_{i=1}^{n} \hat{\ell}_{i,nT}}{n} \tag{4}$$

For the critical value $c_p(n)$, rather than use the traditional critical value from student-t distribution $\Phi^{-1}(1-\frac{P}{2})$, we use:

$$c_p(n) = \Phi^{-1}(1 - \frac{p}{2n^{\delta}}) \tag{5}$$

Suggested by Bailey, Pesaran, and Smith (2019), here, $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution function of a standard normal distribution, P is the size of the test, and δ is a non-negative value represent the critical value exponent. In the scenario of cross-section unit's dimension excess the time observation's dimension, this critical value estimation has been proved that

This estimated critical value, has been showed that, under both Gaussian and non-Gaussian, can provides a true positive rate tend to unit with probability one, meanwhile the type-one error rate converges to zero with probability one.

After obtain the $\hat{\pi}_{nT}$, we can use the following formula provided by Bailey et al. (2020) to estimate our strength indicator α_j:

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

Whenever we have $\hat{\pi}_{nT}$, the estimated $\hat{\alpha}$ will be equal to zero. From the estimation, we can find out that $\hat{\alpha} \in [0,1]$

2 Monte Carlo Design

2.1 Design

In order to study the limited sample property of factor strength α_j, we designed a Monte Carlo simulation. Through the simulation, we compare the property of the factor strength in different settings. Since we will apply the factor strength under the scenario of CAPM model, we consider the following data generating process (DGP): a multi-factor CAPM model.

$$x_{it} = q_1(r_{mt} - r_f) + q_2(\sum_{j=1}^{k} \beta_{ij} f_{jt}) + \varepsilon_{it}$$

In the simulation, we consider a dataset has $i = 1,2,\ldots,n$ different cross-section units, with $t = 1,2,\ldots,T$ different observations. x_{it} is the cross-section return of different asset. f_{jt} represents different risk factors, and the corresponding β_{ij} are the factor loadings. We use $r_{mt} - r_{ft}$ to denotes the market factor. The r_{mt} is the average market return and r_{ft}represent the risk free return. By assumption, the market factor will has strength equals to one all the time, so we consider the market factor as factor f_m which has strength $\alpha_m = 1$. ε_{it} is the stochastic error term. Therefore, the simulation model can be simplified as:

$$x_{it} = q_1(f_{mt}) + q_2(\sum_{j=1}^{k} \beta_{ij} f_{jt}) + \varepsilon_{it}$$

$q_1(\cdot)$ and $q_2(\cdot)$ are two different functions represent the unknown mechanism of market factor and other risk factors in pricing asset risk. In the classical CAPM model and it's multi-factor extensions, for example the three factor model introduced by Fama and French (1992), both q_1 and q_2 are linear.

For each factor, we assume they follow a multinomial distribution with mean zero and a $k \times k$ variance-covariance matrix Σ.

$$\mathbf{f_t} = \begin{pmatrix} f_{i,t} \\ f_{2,t} \\ \vdots \\ f_{k,t} \end{pmatrix} \sim MVN(\mathbf{0}, \Sigma) \quad \Sigma := \begin{pmatrix} \sigma_{f_1}^2, & \rho_{12}\sigma_{f1}\sigma_{f2} & \cdots & \rho_{1k}\sigma_{f1}\sigma_{fk} \\ \rho_{12}\sigma_{f2}\sigma_{f1}, & \sigma_{f2}^2 & \cdots & \rho_{2k}\sigma_{f2}\sigma_{fk} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1k}\sigma_{fk}\sigma_{f1}, & \rho_{k2}\sigma_{fk}\sigma_{f2} & \cdots & \sigma_{fk}^2 \end{pmatrix}$$

The diagonal of matrix Σ indicates the variance of each factor, and the rest represent the correlation among all k factors.

2.2 Baseline Experiment

Follow the general model above, we assume both $q_1(\cdot)$ and $q_2(\cdot)$ are linear function:

$$q_1(f_{mt}) = a_{it} + \beta_{im} f_{mt}$$

$$q_2(\sum_{j=1}^{k} \beta_{ij} f_{jt}) = \sum_{j=1}^{k} \beta_{ij} f_{jt}$$

Therefore, if we include the market factor with other risk factors together, the model can be simplified as:

$$x_{it} = a_{it} + \sum_{j=1}^{k+1} \beta_{ij} f_{jt} + \varepsilon_{it} \tag{6}$$

And in this first baseline experiment, we will use the single factor model as:

$$x_{it} = a_{it} + \beta_{i1} f_{1t} + \varepsilon_{it} \tag{7}$$

Through the simulations, we will control the underlying true strength of factor

To generate factor loadings and asset's return, we first generate the constant term a_{it} which has a uniform distribution from -0.5 to 0.5, $a_{it} \sim U[-0.5, 0.5]$. Then, in this baseline design, we assume the error term follow a standard normal distribution $\varepsilon_{it} \sim N(0,1)$. Next, we set up the true factor strength α. Through the whole simulation, we will assign the strength with different value $\alpha = \{0.5, 0.7, 0.9, 1\}$, and since in this baseline design we only contain one factor, the only factor's strength will be selected from the above set. After having the factor strength, we can calculate for each factor, how many loadings should be different from zero. From the section (1.1), we assume that for any factor with strength α_j, the factor is supposed to generate $[n^{\alpha_j}]$ non-zero factor loadings, and $n - [n^{\alpha_j}]$ zero loadings. Therefore, we can calculate the $n - [n^{\alpha_j}]$.

From the previous section, we assume factors will follow a multinomial standard distribution with mean zero and variance Σ. This means that for each factors, they should follow a normal distribution. In this baseline design, we only contain one factor, and this factor will generate form standard error distribution.

After that, we will generate the factor loadings from a uniform distribution. To make sure every factor loading is sufficiently larger than 0, we set the expected value of those loadings $\mu_\beta = 0.71$,

$\beta_{i1} \sim IIDU(\mu_\beta - 0.2, \mu_\beta + 0.2)$. Then we randomly assign $n - [n^\alpha]$ factor loadings as zero, to reflect the fact that only $[n^\alpha]$ factor loadings are non-zero.

For this experiment, we construct the hypothesis test base on the null hypothesis $H_O : \beta_{i1} = 0$ against the alternative hypothesis $H_1 : \beta_{i1} \neq 0$. The test statistic and critical value are from equation (3) and equation (5). We consider two-sided tests, with size 0.05. Therefore, the corresponding critical value for such t-test will be $\delta = 1.96$

After generate constant term, factor, factor loading, and the error term, we can calculate the simulated asset's return by using the equation (7). With the return and factors, we can re-calculate the factors loading and use the estimation method discussed in section 1.2.

2.3 Two factor experiment

Follow the similar idea as baseline design, we can easily extend the DGP into multi-factor form. We derive a two-factor model from the model (6).

$$x_{it} = a_{it} + \beta_{i1} f_{1t} + \beta_{i2} f_{2t} + \varepsilon_{it} \tag{8}$$

Here $\mathbf{f}_t = (f_{1t}, f_{2t})'$ are two different factors generate from multivariate normal distribution with mean zero and variance Σ. In this simulation, we assume two factors are independent with each other and both of them have variance equals to one, the variance-covariance matrix Σ of factors $\mathbf{f_t}$ will be:

$$\Sigma_{\mathbf{f_t}} := \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Besides, for the factor f_{1t}, we assign it as the market factor, which indicates that the factor strength α_1 will be unit. And all factor loading generates from this factor will be different from zero. For the rest of the variables, we follow the same procedure as the baseline experiment.

In purpose of Monte Carlo Simulation, we consider the different combinations of T and n with $T = \{120, 240, 360\}$, $n = \{100, 300, 500\}$. The market factor, if included in the experiment, will have strength $\alpha_m = 1$ all the time, and the strength of the other factor will be $\alpha_x = \{0.5, 0.7, 0.9, 1\}$. For every setting, we will replicate 500 times independently, all the constant a_{it} and loading β_i will be re-generated for each replication. To exam the goodness of estimation, we calculate the bias between our true underneath factor strength α and the estimated strength $\hat{\alpha}$ as $bias = |\alpha - \hat{\alpha}|$. We also use the bias to calculate the Mean Square Error (MSE). To calculate the MSE, we will collect the bias for each replication, and then use the formula:

$$MSE = \frac{1}{n}\sum_{i=1}^{500}(bias_i)^2$$

2.4 Monte Carlo Discoveries

We report the results in Table (1) and Table (2) for baseline experiment and two-factor experiment respectively. The Table (1) shows the bias and MSE for different α and different (n, T) combinations under the single factor setting. Because the Table (2) is the result of two factor Table (2) shows the bias and MSE for different α_2 with $\alpha_1 = 1$ under different (n, T) combinations. The two tables shows very similar results. The estimation method we applied tend to over-estimate the strength when the true strength is relatively weak. The bias is around 0.2 when the true underlying factor strength is 0.5. Such bias, however, decrease gradually with α rise. When the strength increase to 0.7, the bias will decrease about 0.1 unit. And when the true factor strength is 1,the strongest it can be, we find that the bias and MSE are all converge to zero under all sample size and time combinations.

References

Bailey, N., Kapetanios, G., & Pesaran, M. H. (2016, 9). Exponent of cross-sectional dependence: Estimation and inference. *Journal of Applied Econometrics*, *31*, 929-960. Retrieved from `http://doi.wiley.com/10.1002/jae.2476` doi: 10.1002/jae.2476

Bailey, N., Kapetanios, G., & Pesaran, M. H. (2019, 9). Exponent of cross-sectional dependence for residuals. *Sankhya B*, *81*, 46-102. doi: 10.1007/s13571-019-00196-9

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

Bailey, N., Pesaran, M. H., & Smith, L. V. (2019, 2). A multiple testing approach to the regularisation of large sample correlation matrices. *Journal of Econometrics*, *208*, 507-534. doi: 10.1016/j.jeconom.2018.10.006

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

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*.

A Simulation Result Table

Table 1: Simulation result of single factor model

	Single Factor					
	Bias			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

This table shows the result of one risk factor model. We simulated scenarios of factor strength equals to 0.5, 0.7, 0.9, and 1 with different time, assets size combination. The replication times is 500

Table 2: Simulation result of two factor model

	Two Factor					
	Bias			MSE		
$\alpha_2 = 0.5$, $\alpha_1 = 1.0$						
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
$\alpha_2 = 0.7$, $\alpha_1 = 1.0$						
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
$\alpha_2 = 0.9$, $\alpha_1 = 1.0$						
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
$\alpha_2 = 1.0$, $\alpha_1 = 1.0$						
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

This table shows the result of two factor model, with one market factor and one risk factor. We simulated scenarios of factor strength equals to 0.5, 0.7, 0.9, and 1 with different time, assets size combination. The replication times is 500