Factor Selection and Factor Strength

An Application to U.S. Stock Market Return

Update :August 13, 2020

Version: 0.2.2

Zhiyuan Jiang

I.D:28710967

Supervisor : Dr Natalia Bailey

Dr David Frazier

August 13, 2020

Contents

1 Introduction and Motivation **3**

2 Related Literature **4**

3 Factor Strength **6**

3.1 Definition 6

3.2 Estimation Under single factor setting 7

3.3 Estimation Under Multi-Factor Setting 9

4 Monte Carlo Design **10**

4.1 Design 10

4.2 Experiment Setting 11

4.3 Monte Carlo Results 12

5 Elastic Net **14**

6 Empirical Application **14**

6.1 Data 15

6.2 Factor Strength Analysis 16

6.2.1 Regression model setting 16

6.2.2 Factor Strength Findings 16

6.3 Elastic Net Application 20

7 Conclusion **20**

References **21**

A Simulation Result **25**

B Empirical Application Result **28**

B.1 Factor Strength Estimation Table 28

B.2 Strength Comparison Figures 37

List of Figures

1 Strength Comparison . . . 37

2 Thirty Year Decompose Comparison . . . 43

List of Tables

1 Data Set Dimensions . . . 15

2 Market factor strength estimation . . . 17

3 Proportion of Strength (Excluded Market Factor) . . . 17

4 Selected Risk Factor with Strength: top 15 factors from each data set and three well know factors. . . . 18

5 Simulation result for single factor setting . . . 25

6 Simulation result for double factors setting (no correlation) . . . 26

7 Simulation result for double factors setting (weak correlation) . . . 27

8 Comparison table of estimated factor strength on three different data sets, from strong to weak . . . 28

9 Decompose the thirty year data into three ten year subset, estimated the factor strength base on those three data set separately. Rank the result base on the factor strength, from strong to weak. . . . 33

1 Introduction and Motivation

Capital Asset Pricing Model (CAPM) (Sharpe, 1964; Lintner, 1965; Black, 1972) introduces a risk pricing paradigm. By incorporating factors, the model divides an asset's risk into two parts: systematic risk and asset specified idiosyncratic risk. In general, the market factor proxies for the systematic risk, and different risk factors price the idiosyncratic risk. Researches (see Fama and French (1992), Carhart (1997), Kelly, Pruitt, and Su (2019)) have shown that adding different risk factors into the CAPM model can enhance the ability of pricing risk. Because of this, identifying risk factors has become an important topic in finance. Numerous researchers have contributed to this field, and the direct result is an explosive growth of factors. Harvey and Liu (2019) have documented and categorised over 500 factors from papers published in the top financial and economic journals, and they find the growth of new factors has sped up since 2008.

But we should notice that for a factor, it may not able to capture the risk-return relationship for every asset. Therefore, Pesaran and Smith (2019) introduced a new criterion for assessing the significance of each factor, which they call it factor strength. In general, if a factor can generate coefficients, or refer it as loadings in financial literatures, significantly different from zero for all assets, then we call such a factor strong factor. And the less amounts of significant loadings a factor can generate, the weaker the strength the factor has.

In his 2011 president address Cochrane emphasized the importance of finding factors which can provide independent information about average return and risk. With regard to this, many scholars applied various methods to find such factors. For instance, Harvey and Liu (2017) provided a bootstrap method to adjust the threshold of factor loading's significant test, trying to exclude some falsely significant factor caused by multiple-test problem. Some other scholars are using machine learning methods to reduce the potential candidates. One stream of them has used a shrinkage and subset selection method called Lasso (Tibshirani, 1996) and it's variations to find suitable factors. One example of such an application is made by Rapach, Strauss, and Zhou (2013). They applied the Lasso regression, trying to find some characteristics from a large group to predict the global stock market's return.

But an additional challenge is that factors, especially in the high-dimension, are correlated. Kozak, Nagel, and Santosh (2020) point out that when facing a group of correlated factors, Lasso will only pick several highly correlated factors, seemly at random, and then ignore the other and shrink them to zero. In other words, Lasso fails to handle the issue of correlated factors appropriately.

Therefore, the main empirical question in this project is: how to select useful factors from a large group of possibly highly correlated candidates. We address this question from two different prospects.

From one side, we employ the idea of factor strength discuss above, trying to use this criterion to select those strong factors. On the other hand, we use another variable selection method called Elastic Net (Zou & Hastie, 2005) to select factors. With regard of the first approach, Bailey, Kapetanios, and Pesaran (2020) provide a consistent estimator for the factor strength, and we will use this method to examine the strength of each candidate factor and filter out any spurious factors. Under the second approach, unlike Lasso, elastic net adds an extra penalty term into the loss function, which makes it suitable to handle the potential correlation variables. This trait makes it suitable for our purpose. We will assess and compare the methods in their selection of risk factors. Additionally, we can also use the factor strength as a standard to reduce the dimension of our candidates factors and then applied the elastic net to conduct further selection.

The rest of the thesis is organized as follows. In section 2, we go through some literatures relate with the CAPM model and methods about factor selection. Then in section 3, we will provide a detailed description of the concept of factor strength and the estimation method. In section 4, we set up a simple Monte Carlo simulation experiment to examine the finite sample properties of the factor strength estimator. We introduce the elastic net in section 5. Section 6 includes the empirical application, where we estimate the strength of potential risk factors to be included in a CAPM model, as well as apply elastic net as a method to select factors.

2 Related Literature

This project is built on contributions to the field of asset pricing. First formulated by Sharpe (1964), Lintner (1965), and Black (1972), the original CAPM model only contains the market factor, which is denoted by the difference between average market return and risk-free return. Fama and French (1992) extend the model into three-factors, which it then extend into four (Carhart, 1997), and five

(Fama & French, 2015). Recent research created a six-factors model and claim it outperforms all other sparse factor models. (Kelly et al., 2019).

In terms of assessing the strength of risk factors, this thesis also relates to papers discussing factors that have no or weak correlation with assets' return under the paradigm of the CAPM model. Kan and Zhang (1999) found that the test-statistic of FM two-stage regression (Fama & MacBeth, 1973) will inflate when incorporating factors which are independent of the cross-section return. Therefore, when factors with no pricing power were added into the model, those factors may have the chance to pass the significant test falsely. Kleibergen and Zhan (2015) found out that even when some factor-return relationship does not exist, the r-square and the t-statistic of the FM two-stage regression would become in favour of the conclusion of such structure presence. Gospodinov, Kan, and Robotti (2017) show how the addition of a spurious factor will distort the statistical inference of parameters. Besides, Anatolyev and Mikusheva (2018) studied the behaviours of the model with the presence of weak factors under asymptotic settings, and they find the regression will lead to an inconsistent risk premia estimation result.

Finally, of interest in this thesis is the large dimension of potential factors. For these reasons, it borrows from researchers that identify useful factors from a group of potential factors. Harvey, Liu, and Zhu (2015) examine over 300 factors published in journals, presents a new multi testing framework to exam the significance of factors. And they claim that a higher hurdle for the t-statistic is necessary when examining the significance of newly proposed factors. Methods like a Bayesian procedure introduced by Barillas and Shanken (2018) were used to compare different factor models. Pukthuanthong, Roll, and Subrahmanyam (2019) defined several criteria for ”genuine risk factor”, and based on those criteria introduced a protocol to examine whether a factor is associated with the risk premium.

Once the factor strength is identified, the thesis will attempt to reconcile empirically the factor selection under machine learning techniques and the factor strength implied by the selection.

Gu, Kelly, and Xiu (2020) elaborate on the advantages of using emerging machine learning algorithms in measuring the equity risk premiums. They obtained a higher predictive accuracy in measuring risk premium, and demonstrated large economics gains using investment strategy base on the machine learning forecast. Various machine learning algorithms have been adopted when selecting factors for the factor model, especially in recent years. Lettau and Pelger (2020) apply Principle Components Analysis when investigating the latent factor of the model. Lasso is a popular algorithms for factor selections, because of its ability of selecting features. Feng, Giglio, and Xiu (2019) used the double-selected Lasso method (Belloni, Chernozhukov, & Hansen, 2014), and a grouped lasso method (Huang, Horowitz, & Wei, 2010) is used by Freyberger, Neuhierl, and Weber (2020) when picking factors from a group of candidates. Kozak et al. (2020) arguing that the sparse factor model is ultimately futile by using a Bayesian-based method. They constructed their estimator similar to the ridge regressor, but instead of putting the penalty on the sum of squared of factor coefficients, they impose the penalty base on the maximum squared Sharpe ration implied by the factor model. They also augmented their Bayesian based estimator with extra L^1, created a method, similar but different to the elastic net algorithm which will be employed by our project. (**Still not very good**)

3 Factor Strength

The concept of factor strength employed in this project comes from Bailey et al. (2020), and it was first introduced by Bailey, Kapetanios, and Pesaran (2016). They defined the strength of factor from the prospect of the cross-section dependences of a large panel and connect it to the pervasiveness of the factor, which is captured by the factor loadings. In a separate paper, Bailey, Pesaran, and Smith (2019) extended the method by loosening some restrictions and proved that their estimation can also be applied on the residuals of regression result. Here, we focus on the case of observed factor, and use the method of Bailey et al. (2020) in this project.

3.1 Definition

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

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

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. 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 relates to how many non-zero loadings correspond to a factor. More precisely, for a factor f_{jt} with n different factor loading β_{ij}, we assume that:

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

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

The α_j represents the strength of factor f_{jt} and $\alpha_j \in [0, 1]$. If a 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 the integral operator, which will only take the integral part of the 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 a strong factor. And if we have factor strength $\alpha = 0$, it means that the factor has all factor loadings equal to zero, and we will describe such factor as a weak factor (Bailey et al., 2016). For any factor with strength in [0.5, 1], we will refer such factor as semi-strong factor. In general term, the more non-zero loading a factor has, the stronger the factor's strength is.

3.2 Estimation Under single factor setting

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

To begin with, we consider a single-factor model with the only factor named f_t. β_i is the factor loading of unit i. v_{it} is the stochastic error term.

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

Assume we have n different units and T observations for each unit: $i = 1, 2, 3, \cdots, n$ and $t = 1, 2, 3, \cdots T$. Running the OLS time-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 estimated factor loading of the unit i: $\hat{\beta}_{iT}$, we can construct a t-test to examine its significance under the null hypothesis of the loading is zero. The t-test statistic will be $t_{iT} = \frac{\hat{\beta}_{iT} - 0}{\hat{\sigma}_{iT}}$. Empirically, we calculate the t-statistic of $\hat{\beta}_i$ using:

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

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 can then define an indicator function as: $\ell_{i,n} := \mathbf{1}[|\beta_i| > 0]$. If the factor loading is non-zero, $\ell_{i,n} = 1$. In practice, we use the $\hat{\ell}_{i,nT} := \mathbf{1}[|t_{it}| > c_p(n)]$. Here, if the t-statistic t_{iT} is greater than critical value $c_p(n)$, $\hat{\ell}_{i,n} = 1$, otherwise $\hat{\ell}_{i,n} = 0$. In other words, we are counting how many $\hat{\beta}_{iT}$ is significant. With the indicator function, we then define $\hat{\pi}_{nT}$ as the fraction of significant factor loading amount to the total factor loadings:

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

In term of 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 et al. (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. Adopting this adjusted value helps to tackle the problem of multiple-testing.

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

$$\hat{\alpha} = \begin{cases} 1 + \frac{\ln(\hat{\pi}_{nT})}{\ln n} & \text{if } \hat{\pi}_{nT} > 0, \\ 0, & \text{if } \hat{\pi}_{nT} = 0. \end{cases} \tag{6}$$

When we have the $\hat{\pi}_{nT} = 0$, it means that none of the factor loadings are significantly different from zero, therefore the estimated $\hat{\alpha}$ will be equal to zero. From the estimation, we can find out that $\hat{\alpha} \in [0, 1]$.

3.3 Estimation Under Multi-Factor Setting

This estimation can also be extended into a multi-factor set up. Consider the following multi-factor model:

$$x_{it} = a_i + \sum_{j=1}^{k} \beta_{ij} f_{jt} + v_{it} = a_i + \boldsymbol{\beta_i}' \mathbf{f_t} + v_{it}$$

In this set up, we have $i = 1, 2, \cdots, n$ units, $t = 1, 2, \cdots, T$ time observations, and specially, $j = 1, 2, \cdots, k$ different factors. Here $\beta_i = (\beta_{i1}, \beta_{i2}, \cdots, \beta_{ij})'$ and $\mathbf{f}_t = (f_{1t}, f_{2t} \cdots, f_{jt})$. We employed the same strategy as above, after running OLS and obtain the:

$$x_{it} = \hat{a}_{iT} + \hat{\boldsymbol{\beta}}_{\mathbf{i}}' \mathbf{f}_t + \hat{v}_{it}$$

To conduct the significance test, we calculate the t-statistic: $t_{ijT} = \frac{\hat{\beta}_{ijT} - 0}{\hat{\sigma}_{ijT}}$. Empirically, the test statistic can be calculated using:

$$t_{ijT} = \frac{\left(\mathbf{f}_{j\circ}' \mathbf{M}_{F_{-j}} \mathbf{f}_{j\circ}\right)^{-1/2} \left(\mathbf{f}_{j\circ}' \mathbf{M}_{F_{-j}} \mathbf{x}_i\right)}{\hat{\sigma}_{iT}}$$

Here, $\mathbf{f}_{j\circ} = \left(f_{j1}, f_{j2}, \ldots, f_{jT}\right)', \mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iT})', \mathbf{M}_{F_{-j}} = \mathbf{I} - \mathbf{F}_{-j} \left(\mathbf{F}_{-j}' \mathbf{F}_{-j}\right)^{-1} \mathbf{F}_{-j}'$, and $\mathbf{F}_{-j} = \left(\mathbf{f}_{1\circ}, \ldots, \mathbf{f}_{j-1\circ}, \mathbf{f}_{j+1\circ}, \ldots, \mathbf{f}_{m\circ}\right)'$ For the denominator's $\hat{\sigma}_{iT}$, it was from $\hat{\sigma}_{iT}^2 = T^{-1} \sum_{t=1}^{T} \hat{u}_{it}^2$, the $\hat{u}_{it}$ is the residuals of the model. Then, we can use the same critical value from (5). Obtaining the corresponding ratio $\hat{\pi}_{nTj}$ from (4), and use the function:

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

to estimate the factor strength.

4 Monte Carlo Design

4.1 Design

In order to study the finite sample properties of factor strength $\hat{\alpha}_j$, we conduct a Monte Carlo study. Through the simulation, we compare the property of the factor strength in different settings. We set up the experiments to reflect the CAPM model and its extension. For simplicity, we first define $x_{it} := r_{it} - r_{ft}$. r_{it} is the unit's return, and r_{ft} represent the risk-free rate at time t, therefore, the x_{it} is the excess return of unit i at time t. We use $f_{mt} := r_{mt} - r_{ft}$ to denote the market factor. Here r_{mt} is the average market return of hypothetically all assets in the universe. Additionally, we set $q_1(\cdot)$ and $q_2(\cdot)$ as two different functions that 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. Now consider the following data generating process (DGP):

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

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. f_{jt} represents different risk factors, and the corresponding β_{ij} are the factor loadings. We expect the market factor will have strength equal to one all the time, so we consider the market factor has strength $\alpha_m = 1$. ε_{it} is the stochastic error term.

For each factor, we assume they follow a multivariate normal 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 covariance among all k factors.

4.2 Experiment Setting

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

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

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

To start the simulation, we consider a two-factor model:

$$x_{it} = a_i + \beta_{i1} f_{1t} + \beta_{i2} f_{2t} + \varepsilon_{it} \tag{7}$$

The constant term a_i is generated from a uniform distribution, $a_{it} \sim U[-0.5, 0.5]$. For the factor loading β_{i1} and β_{i2}, we first use a uniform distribution $IIDU(\mu_\beta - 0.2, \mu_\beta + 0.2)$ to produce the values. Here we set $\mu_\beta = 0.71$ to make sure every generated loading value is sufficiently larger than 0. Then we randomly assign $n - [n^{\alpha_1}]$ and $n - [n^{\alpha_2}]$ factor loadings as zero. α_1 and α_2 are the true factor strength of f_1 and f_2. In this simulation, we will start the factor strength from 0.7 and increase it gradually till unity with pace 0.05, say $(\alpha_1, \alpha_2) = \{0.7, 0.75, 0.8, \cdots, 1\}$. $[\cdot]$ is the integer operator defined at section (3.2). This step reflects the fact that only $[n_1^\alpha]$ or $[n_2^\alpha]$ factor loadings are non-zero. In terms of the factors, they come from a multinomial distribution $MVN(\mathbf{0}, \Sigma)$, as we discuss before.

Currently, we consider three different experiments set up:

Experiment 1 (single factor, normal error, no correlation) *Set β_{i2} from (7) as 0, the error term ε_{it} and the factor f_{1t} are both standard normal.*

Experiment 2 (two factors, normal error, no correlation) *Both β_{i1} and β_{i2} are non-zero. Error term and both factors are standard normal. The correlation ρ_{12} between f_{1t} and f_{2t} is zero. The factor strength for the first factor $\alpha_1 = 1$ all the time, and α_2 varies.*

Experiment 3 (two factors, normal error, weak correlation) *Both β_{i1} and β_{i2} are non-zero. Error term and both factors are standard normal. The correlation ρ_{12} between f_{1t} and f_{2t} is 0.3. The factor strength for the first factor $\alpha_1 = 1$ all the time, and α_2 varies.*

The factor strength in each experiment is estimated using the method discussed in section (3.2), the size of the significance test is $p = 0.05$, and the critical value exponent σ has been set as 0.5. For each experiment, we calculate the bias, the RMSE and the size of the test to assess the estimation performances. The bias is calculated as the difference between the true factor strength α and the estimated factor strength $\hat{\alpha}$.

$$bias = \alpha - \hat{\alpha}$$

The Root Square Mean Error (RMSE) comes from:

$$RMSE = [\frac{1}{R}\sum_{r=1}^{R}(bias_r)^2]^{1/2}$$

Where the R represents the total number of replication. The size of the test is under the hypothesis that $H_0 : \hat{\alpha}_j = \alpha_j,\ j = 1,2$ against the alternative hypothesis $H_1 : \hat{\alpha}_j \neq \alpha_j,\ j = 1,2$. Here we employed the following test statistic from Bailey et al. (2020).

$$z_{\hat{\alpha}_j:\alpha_j} = \frac{(\ln n)\left(\hat{\alpha}_j - \alpha_j\right) - p\left(n - n^{\hat{\alpha}_j}\right)n^{-\delta-\hat{\alpha}_j}}{\left[p\left(n - n^{\hat{\alpha}_j}\right)n^{-\delta-2\hat{\alpha}_j}\left(1 - \frac{p}{n^{\delta}}\right)\right]^{1/2}} \quad j = 1,2 \tag{8}$$

Define a indicator function $\mathbf{1}(|z_{\hat{\alpha}_j:\alpha_j}| > c|H_0)$. For each replication, if this test statistic is greater than the critical value of standard normal distribution: $c = 1.96$, the indicator function will return value 1, and 0 otherwise. Therefore, we calculate the size of the test base on:

$$size = \frac{\sum_{r=1}^{R}\mathbf{1}(|z_{\hat{\alpha}_j:\alpha_j}| > 1.96|H_0)}{R} \quad j = 1,2, \tag{9}$$

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 will have strength $\alpha_m = 1$ all the time, and the strength of the other factor will be $\alpha_x = \{0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 1\}$. For every setting, we will replicate 2000 times independently, all the constant and variables will be re-generated for each replication.

4.3 Monte Carlo Results

We report the results in Table (5), (6) and (7) in Appendix A.

Table (5) provides the results under the experiment 1. The estimation method we applied tends to over-estimate the strength slightly most of the time when the true strength is relatively weak under the single factor set up. With the strength increasing, the bias will turn to negative, represents an under-estimated results. Such bias, however, vanishes quickly while observation t, unit amount n, and true strength α increase. When we increase the time spam by including more data from the time dimensions, the bias, as well as the RMSE decrease significantly. Also, when including more cross-section unit n into the simulation, the performance of the estimation improves, as shown by the decreased bias and RMSE values. An impressive result is that the gap between estimation and true strength will go to zero when we have $\alpha = 1$, the strongest strength we can have. With the strength approaching unity, both bias and RMSE will converge to zero. We also present the size of the test in the table. The size of the test will not vary too much when the strength increases, so as the unit increases, But we can observe that when observations for each unit increase, in other words, when t increases, the size will shrink dramatically. The size will become smaller than the 0.05 threshold after we extend the t to 240, or empirically speaking, when we included 20 years monthly return data into the estimation. Notice that, from the equation (8), when $\hat{\alpha} = \alpha = 1$, the nominator becomes zero. Therefore, the size will collapse to zero in all settings, so we do not report the size for $\hat{\alpha} = \alpha = 1$

For the two factors scenarios, we obtain similar conclusions in both the no correlation setting and weak correlation setting. The result of no correlation settings is shown in the table (6), and the table (7) shows the result when the correlation between two factors is 0.3. The estimation results improve when increasing either the observations amount t, or the cross-section units amount n. We also have the same unbiased estimation when true factor strength is unity under all unit-time combinations. In some cases, even when the factor strength is relatively weak, we can have unbiased estimation if the n and t are big enough. (see table (7)). However, we should also notice that when t > n, the results of the size of the test in two factors setting are performing similar to the single factor result. The size will shrink with the observation amount t increasing, and when we have t greater than 240, the size will be smaller than 0.05 threshold in all situations.

5 Elastic Net

Elastic net is variable selection model that can be used for factor selection, introduced by Zou and Hastie (2005). Applying the elastic net method to estimate the factor loading β_{ij} requires:

$$\hat{\beta}_{ij} = \arg\min_{\beta_{ij}} \{ \sum_{i=1}^{n} [(r_{it} - r_{ft}) - \beta_{ij} f_{jt}]^2 + \lambda_2 \sum_{i=1}^{n} \beta_{ij}^2 + \lambda_1 \sum_{i=1}^{n} |\beta_{ij}| \} \tag{4}$$

Because the Lasso regression only contains L_1 penalty term $\sum_{i=1}^{n} |\beta_{ij}|$, it will shows no preference when selecting variables when they are highly correlated. So when Lasso regression will either randomly choose factors from highly correlated candidates, or eliminate them together as a whole. Elastic Net, however, by containing L_2 penalty term $\sum_{i=1}^{n} \beta_{ij}^2$, solves this problem. The L_2 penalty term tend to shrink the potential parameters when they does not provide enough explanatory power, but it will not remove redundant factors. Therefore, the elastic net method will shrink those parameters associated with the correlated factors and keep them, or drop them if they are redundant at pricing risk.

To be complete

6 Empirical Application

Researchers and practitioners have been using the CAPM model (Sharpe (1964), Lintner (1965), and Black (1972)) and its multi-factor extension (For example, the three-factor model by Fama and French (1992)) when they are trying to capture the uncertainty of asset's return. The surging of new factors (Harvey & Liu, 2019) provides numerous option to construct the CAPM model, but it also requires users to pick the factors wisely. In this section, we will use two different methods to identify appropriate factors from a group of 146 candidates. First, we utilise the method introduced in section3 to estimates the strength of each factor from the factor group. Then, we use the strength as a criterion to select factors including in the CAPM model. In the second part of this section, we apply the elastic net method, ask the algorithm to pick factors for us. We will compare the factors selected by those two approaches, expected a consistent selection outcome. Through the empirical application, we also found that under certain conditions, we will obtain a generous CAPM model with too many factors, when using factor strength as the criterion. Therefore, additionally, we apply the elastic net method to further shrinkage down factor selection.

6.1 Data

In the empirical application part, we use the monthly returns on U.S. securities as the assets. The companies are selected from Standard Poor (S&P) 500 index component companies.[1] We prepared three data sets for different time spans: 10 years (January 2008 to December 2017, T = 120), 20 years (January 1998 to December 2017, T = 240), and 30 years (January 1989 to December 2017, T = 360). The initial data set contains 505 companies, but because of the components companies of the index are constantly changing, bankrupt companies will be moved out, and new companies will be added in. Also, some companies do not have enough observations. Therefore, for each of the datasets, the number of companies (n) is different, the dimensions of the data set are showing in the table (1) below.

Table 1: Data Set Dimensions

	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

For the risk-free rate, we use the one-month U.S. treasury bill return.[2] For company i, we calculate the companies return at month t (r_{it}) using the following formula:

$$r_{it} = \frac{p_{it} - p_{it-1}}{p_{it-1}} \times 100$$

and calculate the excess return $x_{it} = r_{it} - r_{ft}$. Here the p_{it} and p_{it-1} are the company's close stock price on the first trading day of month t and t-1. The price is adjusted for the dividends and splits.[3]

Concerning the factors, we use 145 different risk factors from Feng, Giglio, and Xiu (2020). The factor set also includes the market factor, represented by the difference between the average market return and risk-free return. The average market return is a weighted average return of all

[1] The companies return data was obtained from the Global Finance Data: http://www.globalfinancialdata.com/, Osiris: https://www.bvdinfo.com/en-gb/our-products/data/international/osiris, and Yahoo Finance: https://finance.yahoo.com/.

[2] The risk free rate was from the Kenneth R. French website: http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/

[3] The data is adjusted base on the Central for Research in Security Price (CRSP) method.

stocks in the U.S. market, incorporated by CSRP. Each factor contains observations from January 1988 to December 2017.

6.2 Factor Strength Analysis

6.2.1 Regression model setting

For the first part of the empirical application, we estimate the factor strength using the method discussed in section 3. More precisely, we set the regression models based on section 3.3.

$$x_{it} = a_i + \beta_{im}(r_{mt} - r_{ft}) + v_{it}$$

$$x_{it} = a_i + \beta_{im}(r_{mt} - r_{ft}) + \beta_{ij} f_{jt} + v_{it}$$

Here x_{it} is the excess return of asset i at time t, which is pre-defined in section 6.1. $r_{mt} - r_{ft}$ represents the market factor, calculated by the difference between average market return and risk-free return at the same time t. f_{jt} is the value of j^{th} risk factor at time t. Here $j = 1, 2, 3, \cdots 145$. β_{mt} and β_{ij} are the factor loadings for market factor and risk factor, respectively.

We use two different regressions in the purpose of estimating the strength under the single factor setting and the two factors setting. However, due to the potential correlations among factors, we will only focus the market factor strength when using the first single factor regression.

6.2.2 Factor Strength Findings

The complete set of results of factor strength estimation is presented in the appendix B.1 and B.2. We estimated the factors' strength using three different data sets discussed in the 6.1, and rank those strength from strong to weak, alongside the market factor strength, in the table (**??**).

We first look at the market factor strength under the single-factor CAPM setting. (see table 2)

As we expected, the estimated strength of market factor under all three scenarios shows consistently strong results. All three strengths are close to unity, which indicates that the market factor can generate significant factor loading almost all time for every asset. Although the value is close to one, we still notice that the strength will increase slightly with the time span extended. This might indicate that for the security returns, from the long run, it will more closely mimic the behaviours

Table 2: Market factor strength estimation

	Ten Year Data	Twenty Year Data	Thirty Year Data
Market Factor Strength (Single Factor Setting)	0.988	0.990	0.995
Average Market Factor Strength (Double Factors Setting)	0.987	0.957	0.903

of the market than the short run.

Then, we turn to the double factor CAPM setting. We found that for different data sets, the factor strength estimation results are varying. The strongest factor is the market factor for all three data sets. In the ten year data, on average, the market factor has strength 0.987. However, with the observation T increase, the strength of market factor decrease. In twenty-year data set result, the market factor has 0.957, while in the thirty-year, the strength is only 0.903. But, comparing with other factors, the market factor is always the strongest factor.

When looking at other factors, the ten-year data set in general provides a significantly weaker result, compares with the other two data sets results. Except for the market factor, no other factors from the ten-years result show strength above 0.8. The strongest factor besides the market factor is the beta factor which has strength around 0.75. In contrast, the strongest risk factor (factor other than market factor) in the twenty-year data set is the ndp (net debt-to-price), which has strength 0.904. In the thirty-year scenario, the salecash (sales to cash) is the strongest with strength 0.857.

Table 3: Proportion of Strength (Excluded Market Factor)

Strength Level	10 Year Data Proportion	20 Year Data Proportion	30 Year Data Proportion
[0.9, 1]	0%	2.07%	0%
[0.85, 0.9)	0%	24.1%	4.14%
[0.8, 0.85)	0%	16.6%	27.6%
[0.75, 0.8)	0%	8.28%	12.4%
[0.7, 0.75)	7.59%	11.7%	9.66%
[0.65, 07)	15.9%	5.52%	15.9%
[0.6, 0.65)	17.9%	8.28%	5.52%
[0.55, 0.6)	13.1%	8.97%	5.52%
[0.5, 0.55)	8.97%	2.76%	4.83%
[0, 0.5)	36.6%	11.7%	14.5%

When comparing the proportion of factors with strengths falling in different intervals between 0 and 1 (see table (3)), we can find that when using 0.8 as a threshold, there are over forty per

Table 4: Selected Risk Factor with Strength: top 15 factors from each data set and three well know factors.

Ten Year			Twenty Yera			Thirty Year		
Rank	Factor	Strength	Rank	Factor	Strength	Rank	Factor	Strength
1	beta	0.749	1	ndp	0.904	1	salecash	0.857
2	baspread	0.730	2	salecash	0.902	2	ndp	0.852
3	turn	0.728	3	quick	0.901	3	quick	0.851
4	zerotrade	0.725	4	dy	0.897	4	age	0.851
5	idiovol	0.723	5	lev	0.897	5	roavol	0.850
6	retvol	0.721	6	cash	0.897	6	ep	0.849
7	std_turn	0.719	7	zs	0.896	7	depr	0.848
8	HML_Devil	0.719	8	cp	0.894	8	cash	0.847
9	maret	0.715	9	roavol	0.894	9	rds	0.843
10	roavol	0.713	10	age	0.894	10	currat	0.840
11	age	0.703	11	cfp	0.893	11	chcsho	0.840
12	sp	0.699	12	op	0.893	12	zs	0.839
13	ala	0.699	13	nop	0.893	13	nop	0.839
14	ndp	0.686	14	ebp	0.893	14	dy	0.838
15	orgcap	0.686	15	ep	0.891	15	lev	0.838
20	UMD	0.678	28	HML	0.874	38	HML	0.811
24	HML	0.672	76	SMB	0.745	69	SMB	0.721
87	SMB	0.512	88	UMD	0.703	95	UMD	0.672

cent factors in the twenty-year result exceeds this threshold, and such percentage for the thirty-year results is 31%. In ten year results, the number is zero. We also find that nearly 40% of factors from the ten-year dataset show strength less than 0.5, which is almost three times higher than the twenty and thirty years proportion.

Another important finding is that from the twenty-year data set, we obtained three factors: ndq (Net debt-to-price, $\hat{\alpha}$ = 0.904), salecash (sales to cash, $\hat{\alpha}$ = 0.902), and quick (quick ratio, $\hat{\alpha}$ = 0.901) has strength greater than 0.9. We would expect when applying the elastic net method with the twenty-year data set, those three factors with the market factors would be selected.

When looking at the ranking, we found that there are three factors entering the top 15 factor list in all three data sets results. The roavol (Earnings volatility, 10th of ten-year result, 9th of twenty-year result, 5th of thirty-year result), age (Years since first Compustat coverage, 11th of ten-year result, 10th of twenty-year result, 4th of thirty-year result), and ndp (net debt-to-price, 14th of ten-year result, 1st of twenty-year result, 2nd of thirty-year result). This might indicates a persistent risk pricing ability of these three factors exist, even with the changes of the data set's dimensions.

We also focus on some well-known factors, namely the Fama-French size factor (Small Minus Big SMB), Fama-French Value factor (High Minus Low: HML) (Fama & French, 1992) and the Momentum factor (UMD) (Carhart, 1997). It is surprising that none of these three factors enters the top fifteen list for each data sets. Except for the HML factor from the twenty and thirty-year data set has strength above 0.8, none of the other factors in any data set shows strength higher than 0.75 When using the ten-year data, both UMD and HML has strength around 0.67, and the SMB only has strength 0.512. Results from the twenty-year data set show that HML has strength 0.874, for SMB and UMD the strength are 0.745 and 0.703 respectively. Comparing with the twenty-year results, the thirty-year estimated strength drop slightly, HML decreases to 0.811, SMB is 0.721 and UMD has strength 0.672. Therefore, when using the strength as a criterion, we only select the value factor to incorporate in the CAPM model when having twenty and thirty-year data.

As a second step, in order to see how factor strengths evolve through the time, we decompose the thirty yea-data set into three small subsample. For each subsamples, it contains 242 companies (n = 242). And for each company, we obtained 120 observations (t = 120). The results are present in the table (9) and figure (2).

In general, we can conclude that for most of the factors, their strength gradually increased from the first decade (January 1988 to December 1997) to the second decade (January 1998 to December 2007), and then decreased in the third decade (January 2008 to December 2017). This pattern can also be seen in the figure (2). The drop of factor strength in the third decades can be reconciled with the ten-year data results shows a significantly weaker strength than the results from twenty and thirty years data set.

Overall from the factor strength prospect, we would expect that for different time periods, we will have different candidate factors for the CAPM model. For the ten-year data set, we would expect that only the market factor be useful, and therefore the elastic net method applied latter may only select the market factor. If we use the twenty and thirty-year data, we will have a longer list for potential factors, 62 factors from the twenty-year estimation and 45 from the thirty years has strength greater than 0.8. Hence, we would expect the elastic net to select a less parsimonious model.

In terms of the findings we have above, there are several potential explanations. First, if we consider the structure of our data set, we will find that the longer the time span, the fewer companies are included. This is because the S&P index will adjust the component, remove companies with inadequate behaviours, and add in new companies to reflect the market situation. Hence, those 242 companies in the thirty-year data set can be viewed as survivals after a series of financial and economic crisis. We would expect those companies will have above average performances, such as better profitability and administration, compared with other companies.

Another possible explanation the happening of a series of political and financial unease from the time of late 20 century to 2008. Crisis like the Russia financial crisis in 1998, the bankruptcy of Long Term Capital Management (LTCM) in 2000, the dot com bubble crisis in early 21st century and the Global Financial Crisis (GFC) in 2008 creates market disturbances. Such disturbances, however, provides extra correlations among factors. The extra correlations enable some factors provides additional pricing power risk. But we should also notice that the financial market has been disturbed by those crises so, therefore, some mechanism may no longer working properly during that period. Which means that those crises will also have negative influences on factor when they are capturing the risk-return relationship.

We also need to notice that for some factors, their strength will decrease with time. For instance, the gma (gross profitability) factor and convind (convertible debt indicator) factor (see figure 2) has consecutive strength decrease from the 1987-1997 period to 2007-2017 period. And for most of the factors, their strength will decrease significantly from the 1997-2007 period to 2007-2017 period. Therefore, disqualify some factors as the candidate of the CAPM model when using recent year data is inevitable.

6.3 Elastic Net Application

to be added

7 Conclusion

to be added

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. (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. (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

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. 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

Feng, G., Giglio, S., & Xiu, D. (2020, 6). Taming the factor zoo: A test of new factors. *The Journal of Finance*, *75*, 1327-1370. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.12883 doi: 10.1111/jofi.12883

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

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., & 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

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 Simulation Result

Table 5: Simulation result for single factor setting

	Single Factor								
	Bias × 100			RMSE × 100			Size × 100		
	$\alpha_1 = 0.7$								
n\T	120	240	360	120	240	360	120	240	360
100	0.256	0.265	0.227	0.612	0.623	0.560	7.85	7.7	5.55
300	0.185	0.184	0.184	0.363	0.338	0.335	8.9	4.45	4.5
500	0.107	0.124	0.109	0.259	0.248	0.234	6.9	2.5	1.6
	$\alpha_1 = 0.75$								
100	-0.178	-0.159	-0.168	0.490	0.465	0.450	2.5	0.85	0.4
300	0.154	0.156	0.143	0.281	0.258	0.234	9.4	3.7	3.35
500	0.024	0.033	0.263	0.171	0.155	0.148	7.8	2	1.25
	$\alpha_1 = 0.8$								
100	-0.270	-0.265	-0.258	0.434	0.409	0.411	71.4	72.05	71.45
300	-0.052	-0.044	-0.043	0.183	0.149	0.150	10.15	2.45	2.9
500	0.045	0.068	0.067	0.136	0.126	0.121	16.6	6.4	5.9
	$\alpha_1 = 0.85$								
100	0.053	0.062	0.058	0.253	0.228	0.221	6.05	2.95	2.5
300	-0.012	0.009	-0.001	0.124	0.104	0.095	10.55	1.8	1.15
500	-0.026	-0.007	-0.011	0.096	0.073	0.069	13.25	0.9	0.7
	$\alpha_1 = 0.9$								
100	0.025	0.038	0.360	0.191	0.163	0.157	6.85	2	1.65
300	-0.034	-0.018	-0.020	0.099	0.069	0.068	13.2	0.8	0.9
500	-0.025	-0.001	-0.001	0.072	0.044	0.044	22.3	1.95	1.8
	$\alpha_1 = 0.95$								
100	-0.099	-0.088	-0.090	0.156	0.125	0.126	5.6	0.3	0.55
300	-0.046	-0.025	-0.026	0.083	0.045	0.045	22.5	2.2	2.25
500	-0.030	-0.006	-0.006	0.061	0.026	0.025	33.1	4.4	3.8
	$\alpha_1 = 1$								
100	0	0	0	0	0	0	-	-	-
300	0	0	0	0	0	0	-	-	-
500	0	0	0	0	0	0	-	-	-

Notes: This table shows the result of experiment 1. Factors and error are generate from standard normal distribution. Factor loadings come form uniform distribution $IIDU(\mu_\beta - 0.2, \mu_\beta + 0.2)$, and $\mu_\beta = 0.71$. We keep $[n^{\alpha_j}]$ amount of loadings and assign the rest as zero. For each different time-unit combinations, we replicate 2000 times. For the size of the test, we use a two-tail test, under the hypothesis of $H_0, \hat{\alpha}_j = \alpha_j \ j = 1,2$. Cause under the scenarios of $\alpha = 1$, the size of the test will collapse, therefore the table does not report the sizes for $\alpha_1 = 1$.

Table 6: Simulation result for double factors setting (no correlation)

	Double Factor with correlation $\rho_{12} = 0$								
	Bias $\times$ 100			RMSE $\times$ 100			Size $\times$ 100		
	$\alpha_1 = 1, \alpha_2 = 0.7$								
n\T	120	240	360	120	240	360	120	240	360
100	0.567	0.737	0.628	4.062	3.819	3.799	2.95	1.45	1.85
300	0.512	0.611	0.518	2.398	2.103	1.979	6.25	0.55	0.5
500	-0.149	0.08	-0.019	1.796	1.498	1.443	8	0.2	0.1
	$\alpha_1 = 1, \alpha_2 = 0.75$								
100	-3.051	-3.02	-3.092	4.582	4.245	4.248	2.45	0.1	0.10
300	0.491	-1.035	0.640	1.843	1.460	1.576	7.6	0.8	0.55
500	-0.611	-0.372	-0.393	1.520	1.136	1.125	11.35	0.15	0.1
	$\alpha_1 = 1, \alpha_2 = 0.8$								
100	-3.752	-3.630	-3.581	4.557	4.213	4.210	84.65	85.9	85.25
300	-1.218	-0.331	-1.021	1.812	0.792	1.438	9.35	0.2	0.3
500	-0.022	0.192	0.147	1.047	0.782	0.742	15.35	1.1	1.1
	$\alpha_1 = 1, \alpha_2 = 0.85$								
100	-0.075	0.127	0.088	1.996	1.697	1.606	5.4	1.15	0.95
300	-0.531	-0.406	-0.351	1.097	0.613	0.777	10.8	0.15	0.2
500	-0.647	-0.391	-0.391	1.020	0.643	0.630	19.1	0.15	0
	$\alpha_1 = 1, \alpha_2 = 0.9$								
100	-0.128	0.043	0.025	1.428	1.143	1.118	4.9	0.65	0.7
300	-0.651	-0.334	-0.394	1.002	0.435	0.617	17.1	0.6	0.2
500	-0.434	-0.168	-0.171	0.7435	0.367	0.368	25.2	0.4	0.3
	$\alpha_1 = 1, \alpha_2 = 0.95$								
100	-1.218	-1.043	-1.036	1.603	1.222	1.212	6.65	0.25	0.05
300	-0.611	-0.344	-0.356	0.881	0.435	0.434	23.35	0.6	0.45
500	-0.415	-0.123	-0.134	0.661	0.220	0.216	36.75	1.35	1.1
	$\alpha_1 = 1, \alpha_2 = 1$								
100	0	0	0	0	0	0	-	-	-
300	0	0	0	0	0	0	-	-	-
500	0	0	0	0	0	0	-	-	-

Notes: This table shows the result of experiment 2. Factors and errors are generate from standard normal distribution. Between two factors, we assume they have no correlation. Factor loadings come form uniform distribution $IIDU(\mu_\beta - 0.2, \mu_\beta + 0.2)$, and μ_β is set to 0.71. We keep $[n^{\alpha_j}]$ amount of loadings and assign the rest as zero. For each different time-unit combinations, we replicate 2000 times. For the size of the test, we use a two-tail test, under the hypothesis of $H_0, \hat{\alpha}_j = \alpha_j$ $j = 1, 2$. Cause under the scenarios of $\alpha = 1$, the size of the test will collapse, therefore the table does not report the sizes for $\alpha_1 = \alpha_2 = 1$

Table 7: Simulation result for double factors setting (weak correlation)

	Double Factor with correlation $\rho_{12} = 0.3$								
	Bias $\times$ 100			RMSE $\times$ 100			Size $\times$ 100		
$\alpha_1 = 1, \alpha_2 = 0.7$									
n\T	120	240	360	120	240	360	120	240	360
100	0.038	0.064	0.072	0.421	0.382	0.389	4.6	1.75	1.95
300	0.021	0.058	0.056	0.253	0.206	0.198	9.95	0.9	0.25
500	-0.032	0.006	0	0.201	0.153	0	12.20	0.1	0.05
$\alpha_1 = 1, \alpha_2 = 0.75$									
100	-0.325	-0.313	-0.310	0.488	0.419	0.420	4.75	0.1	0
300	0.028	0.063	0.065	0.253	0.157	0.159	9.95	0.55	0.5
500	-0.082	-0.037	-0.039	0.175	0.114	0.112	19.25	0.25	0.3
$\alpha_1 = 1, \alpha_2 = 0.8$									
100	-0.393	-0.361	-0.368	0.477	0.418	0.421	85.45	85.2	86.4
300	0.029	-0.099	-0.100	0.192	0.145	0.145	12.2	0.65	0.5
500	-0.037	-0.016	0.016	0.129	0.074	0.074	27.8	0.25	1.2
$\alpha_1 = 1, \alpha_2 = 0.85$									
100	-0.027	0.008	0.007	0.234	0.160	0.155	9.3	0.9	0.65
300	-0.147	-0.031	-0.037	0.219	0.079	0.077	16.75	0.3	0.2
500	-0.088	-0.039	-0.039	0.136	0.063	0.062	30.6	0.15	0
$\alpha_1 = 1, \alpha_2 = 0.9$									
100	-0.033	0.003	0.002	0.173	0.111	0.110	9.4	0.6	0.55
300	-0.087	-0.040	-0.041	0.131	0.061	0.061	27.8	0.1	0.05
500	-0.070	-0.017	-0.018	0.111	0.037	0.037	41.15	0.6	0.35
$\alpha_1 = 1, \alpha_2 = 0.95$									
100	-0.134	-0.101	-0.104	0.185	0.122	0.122	10.15	0.1	0.15
300	-0.083	-0.034	-0.034	0.118	0.043	0.044	39.35	0.6	0.6
500	-0.062	-0.013	-0.012	0.937	0.022	0.023	51.8	1.25	2.0
$\alpha_1 = 1, \alpha_2 = 1$									
100	0	0	0	0	0	0	-	-	-
300	0	0	0	0	0	0	-	-	-
500	0	0	0	0	0	0	-	-	-

Notes: This table shows the result of experiment 2. Factors and errors are generate from standard normal distribution. Between two factors, we assume they have correlation $\rho_{12} = 0.3$ Factor loadings come form uniform distribution $IIDU(\mu_\beta - 0.2, \mu_\beta + 0.2)$, and μ_β is set to 0.71. We keep $[n^{\alpha_j}]$ amount of loadings and assign the rest as zero. For each different time-unit combinations, we replicate 2000 times. For the size of the test, we use a two-tail test, under the hypothesis of $H_0, \hat{\alpha}_j = \alpha_j$ $j = 1, 2$. Cause under the scenarios of $\alpha = 1$, the size of the test will collapse, therefore the table does not report the sizes when $\alpha_1 = \alpha_2 = 1$

B Empirical Application Result

B.1 Factor Strength Estimation Table

Table 8: Comparison table of estimated factor strength on three different data sets, from strong to weak

	Ten Year Data			Twenty Year Data			Thirty Year Data		
	Factor	Market Factor Strength	Risk Factor Strength	Factor	Market Factor Strength	Risk Factor Strength	Factor	Market Factor Strength	Risk Factor Strength
1	beta	0.976	0.749	ndp	0.960	0.904	salecash	0.905	0.857
2	baspread	0.980	0.730	salecash	0.958	0.902	ndp	0.905	0.852
3	turn	0.983	0.728	quick	0.958	0.901	quick	0.905	0.851
4	zerotrade	0.983	0.725	dy	0.957	0.897	age	0.905	0.851
5	idiovol	0.981	0.723	lev	0.959	0.897	roavol	0.904	0.850
6	retvol	0.978	0.721	cash	0.958	0.897	ep	0.905	0.849
7	std_turn	0.983	0.719	zs	0.959	0.896	depr	0.905	0.848
8	HML_Devil	0.989	0.719	cp	0.960	0.894	cash	0.905	0.847
9	maxret	0.981	0.715	roavol	0.957	0.894	rds	0.905	0.843
10	roavol	0.985	0.713	age	0.959	0.894	currat	0.905	0.840
11	age	0.989	0.703	cfp	0.960	0.893	chcsho	0.905	0.840
12	sp	0.985	0.699	op	0.958	0.893	zs	0.903	0.839
13	ala	0.986	0.699	nop	0.958	0.893	nop	0.904	0.839
14	ndp	0.987	0.686	ebp	0.959	0.893	dy	0.905	0.838
15	orgcap	0.989	0.686	ep	0.958	0.891	lev	0.903	0.838
16	tang	0.990	0.683	rds	0.958	0.890	cfp	0.905	0.838
17	ebp	0.988	0.683	depr	0.958	0.889	stdacc	0.905	0.837
18	invest	0.986	0.683	sp	0.958	0.888	cp	0.905	0.836
19	dpia	0.986	0.681	currat	0.958	0.887	stdcf	0.905	0.836
20	UMD	0.989	0.678	kz	0.958	0.887	op	0.904	0.835
21	zs	0.986	0.675	chcsho	0.957	0.884	ebp	0.903	0.835
22	grltnoa	0.988	0.675	tang	0.960	0.884	tang	0.904	0.833
23	dy	0.988	0.672	ato	0.958	0.884	kz	0.903	0.831
24	HML	0.987	0.672	stdacc	0.958	0.883	ato	0.904	0.831
25	kz	0.986	0.669	adm	0.958	0.881	ww	0.904	0.827
26	ob_a	0.989	0.669	cashpr	0.959	0.878	std_turn	0.902	0.826
27	BAB	0.989	0.666	stdcf	0.956	0.878	adm	0.904	0.825

Table 8: Comparison table of estimated factor strength on three different data sets, from strong to weak (Cont.)

	Ten Year Data			Twenty Year Data			Thirty Year Data		
	Factor	Market Factor Strength	Risk Factor Strength	Factor	Market Factor Strength	Risk Factor Strength	Factor	Market Factor Strength	Risk Factor Strength
28	op	0.990	0.663	HML	0.958	0.874	idiovol	0.902	0.825
29	realestate_hxz	0.987	0.663	nef	0.956	0.873	maxret	0.902	0.825
30	ol	0.987	0.663	std_turn	0.956	0.870	baspread	0.902	0.820
31	adm	0.988	0.660	idiovol	0.955	0.870	IPO	0.905	0.818
32	lev	0.986	0.657	zerotrade	0.953	0.865	nef	0.902	0.818
33	nxf	0.989	0.651	turn	0.955	0.864	sp	0.903	0.817
34	nop	0.989	0.651	ww	0.959	0.863	turn	0.902	0.813
35	pm	0.986	0.648	maxret	0.956	0.863	retvol	0.902	0.813
36	pchcapx3	0.988	0.644	absacc	0.960	0.859	zerotrade	0.900	0.812
37	nef	0.988	0.644	baspread	0.955	0.854	absacc	0.905	0.812
38	cash	0.989	0.637	hire	0.959	0.851	HML	0.903	0.811
39	QMJ	0.978	0.637	IPO	0.960	0.850	lgr	0.905	0.810
40	rds	0.989	0.634	lgr	0.959	0.850	cashpr	0.903	0.808
41	LIQ_PS	0.988	0.634	nxf	0.956	0.849	dcol	0.905	0.807
42	ato	0.988	0.634	retvol	0.955	0.848	beta	0.900	0.806
43	salerec	0.992	0.630	salerec	0.957	0.847	RMW	0.904	0.806
44	currat	0.989	0.626	RMW	0.957	0.847	hire	0.905	0.805
45	acc	0.989	0.619	beta	0.954	0.846	salerec	0.905	0.803
46	stdcf	0.989	0.619	sin	0.959	0.844	nxf	0.903	0.801
47	HXZ_ROE	0.989	0.619	acc	0.960	0.843	acc	0.904	0.797
48	depr	0.988	0.615	bm_ia	0.960	0.843	dfin	0.902	0.791
49	noa	0.989	0.615	dcol	0.959	0.838	nincr	0.904	0.790
50	cashpr	0.987	0.615	dfin	0.959	0.838	noa	0.902	0.787
51	absacc	0.989	0.615	HML_Devil	0.953	0.838	HML_Devil	0.902	0.781
52	gma	0.987	0.615	HXZ_IA	0.960	0.838	HXZ_IA	0.904	0.780
53	dncl	0.986	0.611	nincr	0.959	0.834	rdm	0.904	0.778
54	ms	0.980	0.611	rna	0.958	0.826	rna	0.904	0.778
55	rna	0.989	0.611	noa	0.957	0.825	rd	0.903	0.774
56	STR	0.987	0.607	herf	0.957	0.824	bm_ia	0.904	0.772
57	rdm	0.988	0.607	rdm	0.958	0.823	sgr	0.904	0.769
58	chcsho	0.987	0.607	sgr	0.958	0.819	ps	0.904	0.769
59	sin	0.987	0.607	dnco	0.959	0.816	sin	0.904	0.769
60	salecash	0.989	0.602	ps	0.957	0.807	realestate_hxz	0.905	0.769

Table 8: Comparison table of estimated factor strength on three different data sets, from strong to weak (Cont.)

	Ten Year Data			Twenty Year Data			Thirty Year Data		
	Factor	Market Factor Strength	Risk Factor Strength	Factor	Market Factor Strength	Risk Factor Strength	Factor	Market Factor Strength	Risk Factor Strength
61	dnco	0.988	0.598	CMA	0.960	0.805	herf	0.902	0.766
62	quick	0.989	0.593	egr_hxz	0.958	0.803	dnco	0.904	0.761
63	stdacc	0.989	0.593	realestate_hxz	0.957	0.798	CMA	0.905	0.759
64	poa	0.988	0.593	gad	0.958	0.788	egr_hxz	0.904	0.750
65	cp	0.988	0.589	rd	0.958	0.787	ob_a	0.903	0.745
66	tb	0.988	0.589	ol	0.954	0.787	ol	0.902	0.741
67	HXZ_IA	0.987	0.584	cinvest_a	0.959	0.784	cinvest_a	0.903	0.739
68	saleinv	0.987	0.579	dolvol	0.960	0.774	gad	0.902	0.723
69	cfp	0.988	0.579	ob_a	0.955	0.764	SMB	0.902	0.721
70	egr	0.987	0.579	ala	0.958	0.762	dolvol	0.904	0.715
71	dnca	0.986	0.579	pchdepr	0.959	0.761	gma	0.902	0.715
72	egr_hxz	0.988	0.579	BAB	0.960	0.757	ala	0.904	0.715
73	os	0.984	0.569	gma	0.955	0.756	cto	0.902	0.710
74	pps	0.983	0.563	pchcapx3	0.957	0.752	aeavol	0.905	0.710
75	cto	0.987	0.563	dnca	0.958	0.747	BAB	0.905	0.710
76	grltnoa_hxz	0.986	0.563	SMB	0.957	0.745	convind	0.904	0.710
77	cei	0.988	0.563	poa	0.957	0.739	tb	0.902	0.708
78	CMA	0.988	0.563	aeavol	0.961	0.737	QMJ	0.903	0.708
79	em	0.989	0.552	tb	0.953	0.732	pricedelay	0.904	0.701
80	ww	0.990	0.546	grltnoa_hxz	0.958	0.730	egr	0.902	0.699
81	std_dolvol	0.987	0.539	cei	0.953	0.730	orgcap	0.902	0.699
82	grcapx	0.986	0.539	indmom	0.956	0.725	pchdepr	0.903	0.696
83	pctacc	0.989	0.539	egr	0.958	0.725	indmom	0.902	0.696
84	ep	0.989	0.533	moms12m	0.957	0.725	dcoa	0.902	0.696
85	pricedelay	0.989	0.533	dsti	0.957	0.723	moms12m	0.903	0.694
86	hire	0.988	0.519	orgcap	0.956	0.715	pchcapx3	0.902	0.691
87	SMB	0.987	0.512	pchcurrat	0.958	0.710	cei	0.902	0.691
88	pchcapx_ia	0.989	0.512	UMD	0.951	0.706	roic	0.902	0.691
89	aeavol	0.988	0.512	dcoa	0.959	0.706	pm	0.903	0.691
90	moms12m	0.987	0.512	roic	0.951	0.703	dnca	0.902	0.689
91	cashdebt	0.984	0.504	QMJ	0.951	0.703	saleinv	0.903	0.686
92	lgr	0.987	0.504	cinvest	0.958	0.701	grltnoa_hxz	0.903	0.683
93	cinvest	0.988	0.496	HXZ_ROE	0.957	0.699	poa	0.903	0.681

Table 8: Comparison table of estimated factor strength on three different data sets, from strong to weak (Cont.)

	Ten Year Data			Twenty Year Data			Thirty Year Data		
	Factor	Market Factor Strength	Risk Factor Strength	Factor	Market Factor Strength	Risk Factor Strength	Factor	Market Factor Strength	Risk Factor Strength
94	herf	0.987	0.496	cto	0.955	0.694	HXZ_ROE	0.905	0.678
95	bm_ia	0.988	0.487	pctacc	0.954	0.694	UMD	0.902	0.672
96	cfp_ia	0.987	0.479	pricedelay	0.958	0.691	pctacc	0.902	0.672
97	cinvest_a	0.989	0.479	pchcapx_ia	0.957	0.681	cinvest	0.903	0.660
98	chmom	0.989	0.469	convind	0.955	0.669	dsti	0.902	0.660
99	RMW	0.987	0.469	cdi	0.958	0.654	em	0.902	0.657
100	sue	0.987	0.459	rsup	0.957	0.651	pchcurrat	0.902	0.654
101	mom36m	0.986	0.459	chtx	0.958	0.644	ms	0.902	0.648
102	indmom	0.987	0.459	invest	0.957	0.644	invest	0.902	0.641
103	dcoa	0.988	0.459	em	0.952	0.644	pchcapx_ia	0.902	0.630
104	etr	0.986	0.448	pm	0.957	0.641	os	0.900	0.623
105	chinv	0.988	0.448	saleinv	0.955	0.637	chtx	0.902	0.623
106	ill	0.988	0.448	ta	0.958	0.634	dpia	0.902	0.623
107	roic	0.986	0.448	dpia	0.957	0.634	cdi	0.903	0.623
108	convind	0.988	0.448	pchquick	0.957	0.626	pps	0.902	0.611
109	sgr	0.988	0.437	os	0.948	0.626	roaq	0.900	0.602
110	IPO	0.989	0.437	ms	0.950	0.619	rs	0.902	0.584
111	dolvol	0.989	0.437	roaq	0.953	0.607	rsup	0.902	0.579
112	dcol	0.987	0.425	grcapx	0.955	0.593	chinv	0.902	0.569
113	nincr	0.989	0.411	pps	0.952	0.589	cfp_ia	0.902	0.563
114	chempia	0.987	0.411	ndf	0.957	0.589	ta	0.903	0.563
115	rs	0.988	0.411	cfp_ia	0.957	0.584	cashdebt	0.900	0.557
116	pchcapx	0.988	0.411	dncl	0.957	0.584	ndf	0.902	0.557
117	chtx	0.988	0.397	pchsale_pchrect	0.955	0.574	grcapx	0.902	0.552
118	ivg	0.988	0.381	mom6m	0.958	0.569	STR	0.902	0.546
119	LTR	0.985	0.364	rs	0.955	0.563	pchcapx	0.902	0.546
120	mom6m	0.987	0.364	pchcapx	0.958	0.563	pchquick	0.902	0.539
121	cdi	0.987	0.364	cashdebt	0.951	0.557	grltnoa	0.902	0.539
122	chatoia	0.987	0.364	pchsaleinv	0.955	0.557	pchsaleinv	0.902	0.519
123	gad	0.985	0.364	chempia	0.958	0.557	dncl	0.902	0.519
124	pchcurrat	0.988	0.297	LIQ_PS	0.956	0.557	ivg	0.902	0.504
125	pchgm_pchsale	0.988	0.297	dwc	0.955	0.546	mom6m	0.902	0.496
126	rd	0.986	0.297	grltnoa	0.956	0.533	chempia	0.902	0.496

Table 8: Comparison table of estimated factor strength on three different data sets, from strong to weak (Cont.)

	Ten Year Data			Twenty Year Data			Thirty Year Data		
	Factor	Market Factor Strength	Risk Factor Strength	Factor	Market Factor Strength	Risk Factor Strength	Factor	Market Factor Strength	Risk Factor Strength
127	dsti	0.989	0.297	STR	0.956	0.526	LIQ_PS	0.902	0.496
128	dfnl	0.987	0.297	dfnl	0.955	0.519	mom36m	0.902	0.479
129	roaq	0.986	0.297	mom36m	0.957	0.496	std_dolvol	0.903	0.459
130	pchdepr	0.988	0.266	std_dolvol	0.955	0.496	pchsale_pchinvt	0.902	0.448
131	dnoa	0.988	0.230	sue	0.956	0.487	pchsale_pchxsga	0.902	0.448
132	ta	0.988	0.230	LTR	0.954	0.487	dwc	0.902	0.448
133	chpmia	0.987	0.230	chmom	0.953	0.479	dfnl	0.902	0.437
134	pchquick	0.987	0.182	pchsale_pchinvt	0.955	0.448	chmom	0.902	0.437
135	dfin	0.988	0.182	chatoia	0.957	0.437	pchsale_pchrect	0.902	0.425
136	rsup	0.988	0.182	pchsale_pchxsga	0.957	0.425	sue	0.902	0.397
137	pchsaleinv	0.988	0.115	lfe	0.956	0.425	LTR	0.902	0.381
138	pchsale_pchinvt	0.988	0.115	chinv	0.956	0.397	pchgm_pchsale	0.902	0.322
139	pchsale_pchrect	0.988	0.115	ivg	0.957	0.397	lfe	0.902	0.297
140	ps	0.990	0.115	pchgm_pchsale	0.957	0.381	ill	0.902	0.297
141	dwc	0.989	0.115	etr	0.955	0.344	dnoa	0.902	0.182
142	pchsale_pchxsga	0.989	0.000	chpmia	0.957	0.344	ear	0.903	0.182
143	lfe	0.988	0.000	ill	0.955	0.266	chatoia	0.902	0.182
144	ndf	0.986	0.000	dnoa	0.955	0.266	chpmia	0.902	0.182
145	ear	0.988	0.000	ear	0.958	0.266	etr	0.902	0.115

Notes: This table presents the estimation results of factors' strength, ordered decreasingly by risk factor strength. For the estimation, we use the method from Section 3.3, with one market factor and one risk factor. The three data set is describe in the section 6.1

Table 9: Decompose the thirty year data into three ten year subset, estimated the factor strength base on those three data set separately. Rank the result base on the factor strength, from strong to weak.

	January 1988 to December 1997		January 1998 to December 2007		January 2008 to December 2017	
Rank	Factor	Strength	Factor	Strength	Factor	Strength
1	herf	0.69	ep	0.83	beta	0.70
2	turn	0.67	roavol	0.83	baspread	0.66
3	saleinv	0.66	nop	0.82	turn	0.66
4	beta	0.66	salecash	0.82	zerotrade	0.66
5	cto	0.66	ndp	0.82	retvol	0.66
6	nef	0.66	dy	0.82	std_turn	0.66
7	zerotrade	0.65	depr	0.82	idiovol	0.66
8	ala	0.65	cp	0.82	roavol	0.65
9	idiovol	0.64	quick	0.82	maxret	0.65
10	ol	0.64	op	0.82	sp	0.63
11	depr	0.64	cash	0.82	age	0.63
12	std_turn	0.64	lev	0.82	dy	0.63
13	gma	0.64	age	0.82	tang	0.63
14	dy	0.63	cfp	0.82	ol	0.62
15	retvol	0.63	ebp	0.82	HML_Devil	0.62
16	baspread	0.63	HML	0.81	op	0.61
17	currat	0.63	zs	0.81	ala	0.61
18	op	0.63	kz	0.81	realestate_hxz	0.61
19	nxf	0.62	currat	0.81	dpia	0.61
20	tang	0.61	rds	0.81	invest	0.61
21	nop	0.61	baspread	0.81	orgcap	0.61
22	maxret	0.61	chcsho	0.81	pm	0.60
23	pm	0.61	beta	0.81	ob_a	0.60
24	orgcap	0.61	sp	0.80	nop	0.60
25	quick	0.61	stdcf	0.80	BAB	0.60
26	SMB	0.61	zerotrade	0.80	grltnoa	0.59
27	sp	0.59	stdacc	0.80	HML	0.59
28	roavol	0.59	maxret	0.80	nef	0.59
29	pricedelay	0.59	retvol	0.80	nxf	0.58
30	aeavol	0.59	std_turn	0.80	UMD	0.58
31	convind	0.59	idiovol	0.80	absacc	0.58
32	bm_ia	0.58	nef	0.80	currat	0.57
33	cash	0.58	turn	0.80	acc	0.57
34	ndp	0.57	ato	0.80	pchcapx3	0.57
35	ivg	0.57	tang	0.79	dncl	0.56
36	hire	0.57	ww	0.79	salerec	0.56
37	egr_hxz	0.56	cashpr	0.79	ndp	0.56
38	roic	0.56	adm	0.79	ebp	0.56
39	HXZ_IA	0.56	IPO	0.78	adm	0.55
40	salerec	0.56	HML_Devil	0.77	LIQ_PS	0.55
41	cp	0.56	RMW	0.77	cash	0.55
42	chinv	0.56	dfin	0.77	QMJ	0.55
43	dcoa	0.56	sin	0.77	rds	0.55
44	dnco	0.56	acc	0.76	stdcf	0.55
45	ebp	0.56	nxf	0.76	lev	0.54
46	cashpr	0.56	bm_ia	0.76	zs	0.54
47	invest	0.56	absacc	0.76	kz	0.54

Table 9: Decompose the thirty year data into three ten year subset, estimated the factor strength base on those three data set separately(cont.)

	January 1988 to December 1997		January 1998 to December 2007		January 2008 to December 2017	
Rank	Factor	Strength	Factor	Strength	Factor	Strength
48	HML_Devil	0.56	salerec	0.75	saleinv	0.53
49	poa	0.55	lgr	0.75	depr	0.53
50	sgr	0.55	dcol	0.75	STR	0.53
51	age	0.55	hire	0.75	ato	0.53
52	dpia	0.55	noa	0.74	poa	0.53
53	cdi	0.55	nincr	0.74	chcsho	0.53
54	em	0.55	dnco	0.74	stdacc	0.53
55	QMJ	0.55	herf	0.73	salecash	0.52
56	salecash	0.54	ala	0.73	cto	0.52
57	HML	0.54	rdm	0.73	noa	0.52
58	egr	0.54	ps	0.71	rna	0.52
59	pchcapx3	0.54	rd	0.71	cashpr	0.52
60	dcol	0.53	HXZ_IA	0.71	gma	0.52
61	acc	0.52	cinvest_a	0.71	HXZ_IA	0.52
62	zs	0.52	BAB	0.71	HXZ_ROE	0.52
63	nincr	0.52	sgr	0.70	quick	0.51
64	rdm	0.52	dolvol	0.70	cp	0.51
65	kz	0.52	SMB	0.70	rdm	0.51
66	rsup	0.52	realestate_hxz	0.70	dnco	0.51
67	pps	0.51	rna	0.69	egr_hxz	0.51
68	grltnoa_hxz	0.51	ndf	0.68	ms	0.51
69	cfp	0.51	CMA	0.68	pps	0.50
70	pctacc	0.51	pchdepr	0.67	egr	0.50
71	ep	0.50	egr_hxz	0.67	ww	0.50
72	lev	0.50	cei	0.67	grltnoa_hxz	0.49
73	chcsho	0.50	poa	0.66	cfp	0.49
74	UMD	0.50	QMJ	0.66	dnca	0.49
75	lgr	0.50	moms12m	0.66	os	0.47
76	CMA	0.50	ol	0.66	sin	0.47
77	dnca	0.49	tb	0.65	pctacc	0.47
78	chmom	0.49	roic	0.64	tb	0.45
79	cei	0.49	ob_a	0.64	grcapx	0.45
80	indmom	0.48	indmom	0.63	cei	0.45
81	dwc	0.48	HXZ_ROE	0.62	hire	0.45
82	realestate_hxz	0.48	gma	0.62	CMA	0.45
83	STR	0.47	aeavol	0.61	SMB	0.44
84	absacc	0.47	orgcap	0.61	pchcapx_ia	0.44
85	RMW	0.45	rsup	0.61	std_dolvol	0.44
86	os	0.44	UMD	0.61	ep	0.42
87	chempia	0.44	grltnoa_hxz	0.60	cfp_ia	0.42
88	rds	0.44	cinvest	0.60	aeavol	0.42
89	ob_a	0.44	pchcapx3	0.60	RMW	0.42
90	ms	0.44	pchcurrat	0.59	pricedelay	0.41
91	ps	0.42	pctacc	0.59	dcoa	0.41
92	grltnoa	0.42	pchcapx_ia	0.59	herf	0.41
93	roaq	0.42	dfnl	0.59	em	0.41
94	BAB	0.42	dnca	0.58	dolvol	0.40
95	grcapx	0.41	gad	0.58	chinv	0.40
96	rs	0.41	mom36m	0.57	cinvest	0.40

Table 9: Decompose the thirty year data into three ten year subset, estimated the factor strength base on those three data set separately(cont.)

	January 1988 to December 1997		January 1998 to December 2007		January 2008 to December 2017	
Rank	Factor	Strength	Factor	Strength	Factor	Strength
97	moms12m	0.41	pchquick	0.57	lgr	0.40
98	chatoia	0.41	cfp_ia	0.57	mom36m	0.38
99	mom36m	0.40	pricedelay	0.56	rs	0.38
100	dolvol	0.40	egr	0.56	moms12m	0.38
101	cinvest	0.40	dsti	0.56	LTR	0.36
102	ww	0.40	convind	0.56	IPO	0.36
103	std_dolvol	0.38	ms	0.56	nincr	0.36
104	stdcf	0.38	cdi	0.55	indmom	0.36
105	chtx	0.36	dcoa	0.55	bm_ia	0.36
106	rd	0.36	dncl	0.55	chempia	0.36
107	ato	0.36	cto	0.54	cinvest_a	0.36
108	HXZ_ROE	0.36	cashdebt	0.53	roic	0.36
109	cashdebt	0.34	ta	0.53	cashdebt	0.34
110	ta	0.34	roaq	0.53	sgr	0.34
111	dfnl	0.34	pchsaleinv	0.52	ill	0.32
112	pchcapx	0.32	pchsale_pchrect	0.52	chmom	0.32
113	stdacc	0.32	chtx	0.50	etr	0.30
114	adm	0.30	em	0.50	ivg	0.30
115	noa	0.30	pchsale_pchinvt	0.50	convind	0.30
116	pchsale_pchinvt	0.27	mom6m	0.48	pchgm_pchsale	0.27
117	cfp_ia	0.27	pm	0.48	chtx	0.27
118	cinvest_a	0.27	pchsale_pchxsga	0.47	dcol	0.27
119	dfin	0.27	os	0.47	cdi	0.27
120	rna	0.27	pchcapx	0.45	sue	0.23
121	sue	0.23	chempia	0.44	mom6m	0.23
122	IPO	0.23	dwc	0.44	dfnl	0.23
123	pchsale_pchxsga	0.23	saleinv	0.42	pchcapx	0.23
124	etr	0.23	LIQ_PS	0.40	chatoia	0.23
125	lfe	0.23	pchgm_pchsale	0.38	pchcurrat	0.18
126	ndf	0.23	pps	0.34	pchdepr	0.18
127	LTR	0.18	rs	0.34	dwc	0.18
128	pchsaleinv	0.18	lfe	0.32	chpmia	0.18
129	mom6m	0.18	LTR	0.30	gad	0.18
130	ill	0.18	std_dolvol	0.30	rd	0.11
131	sin	0.18	chinv	0.27	dnoa	0.11
132	pchsale_pchrect	0.11	chmom	0.27	dfin	0.11
133	LIQ_PS	0.11	grcapx	0.27	dsti	0.11
134	tb	0.11	STR	0.23	rsup	0.11
135	dncl	0.11	ill	0.23	roaq	0.11
136	dsti	0.11	chpmia	0.23	pchquick	0.00
137	ear	0.11	sue	0.18	pchsaleinv	0.00
138	pchcurrat	0.00	grltnoa	0.18	pchsale_pchinvt	0.00
139	pchquick	0.00	dnoa	0.18	pchsale_pchrect	0.00
140	pchdepr	0.00	ear	0.18	pchsale_pchxsga	0.00
141	pchgm_pchsale	0.00	chatoia	0.18	lfe	0.00
142	pchcapx_ia	0.00	etr	0.11	ps	0.00
143	dnoa	0.00	ivg	0.11	ta	0.00
144	chpmia	0.00	dpia	0.00	ndf	0.00

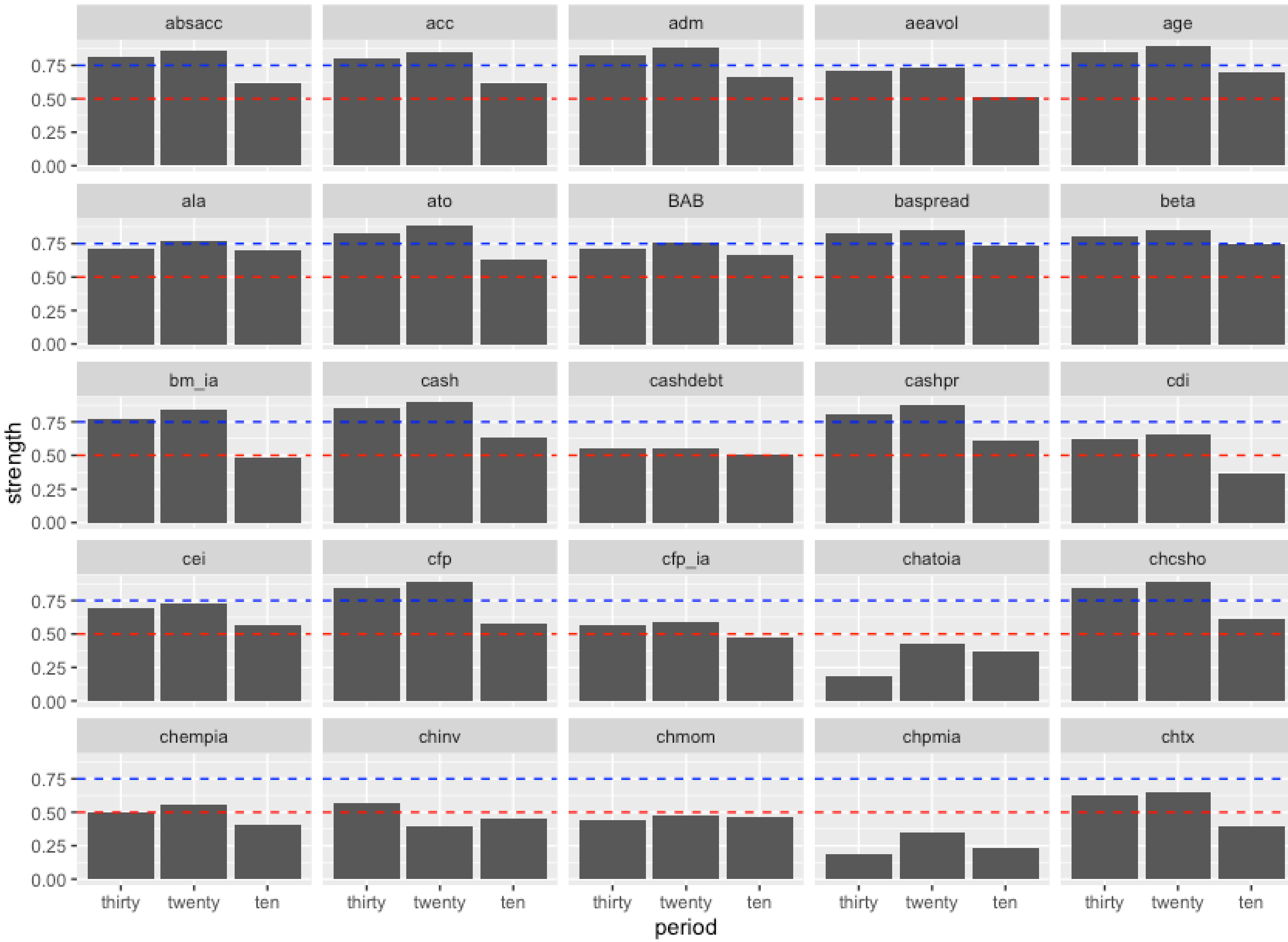
Table 9: Decompose the thirty year data into three ten year subset, estimated the factor strength base on those three data set separately(cont.)

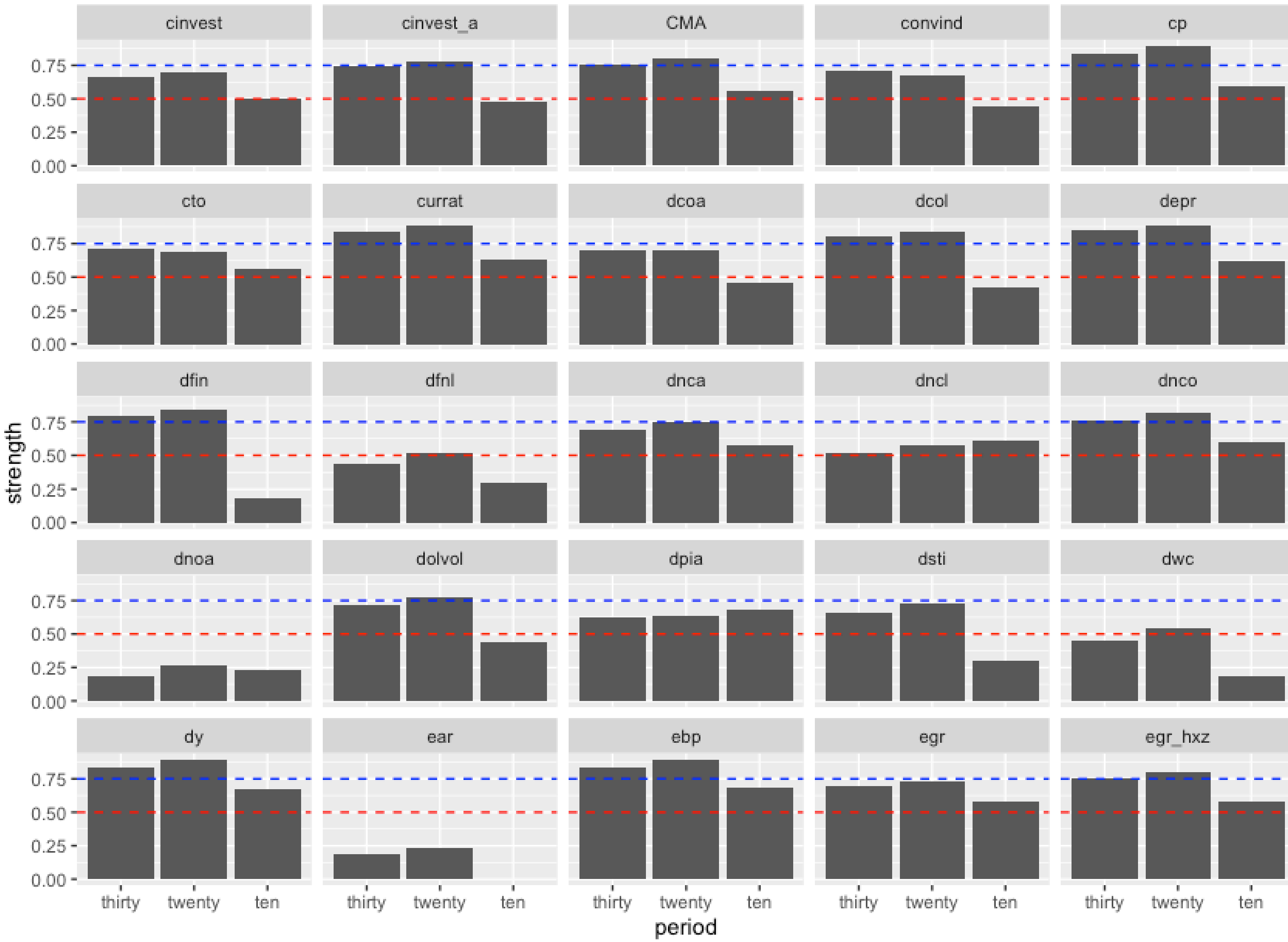
	January 1988 to December 1997		January 1998 to December 2007		January 2008 to December 2017	
Rank	Factor	Strength	Factor	Strength	Factor	Strength
145	gad	0.00	invest	0.00	ear	0.00

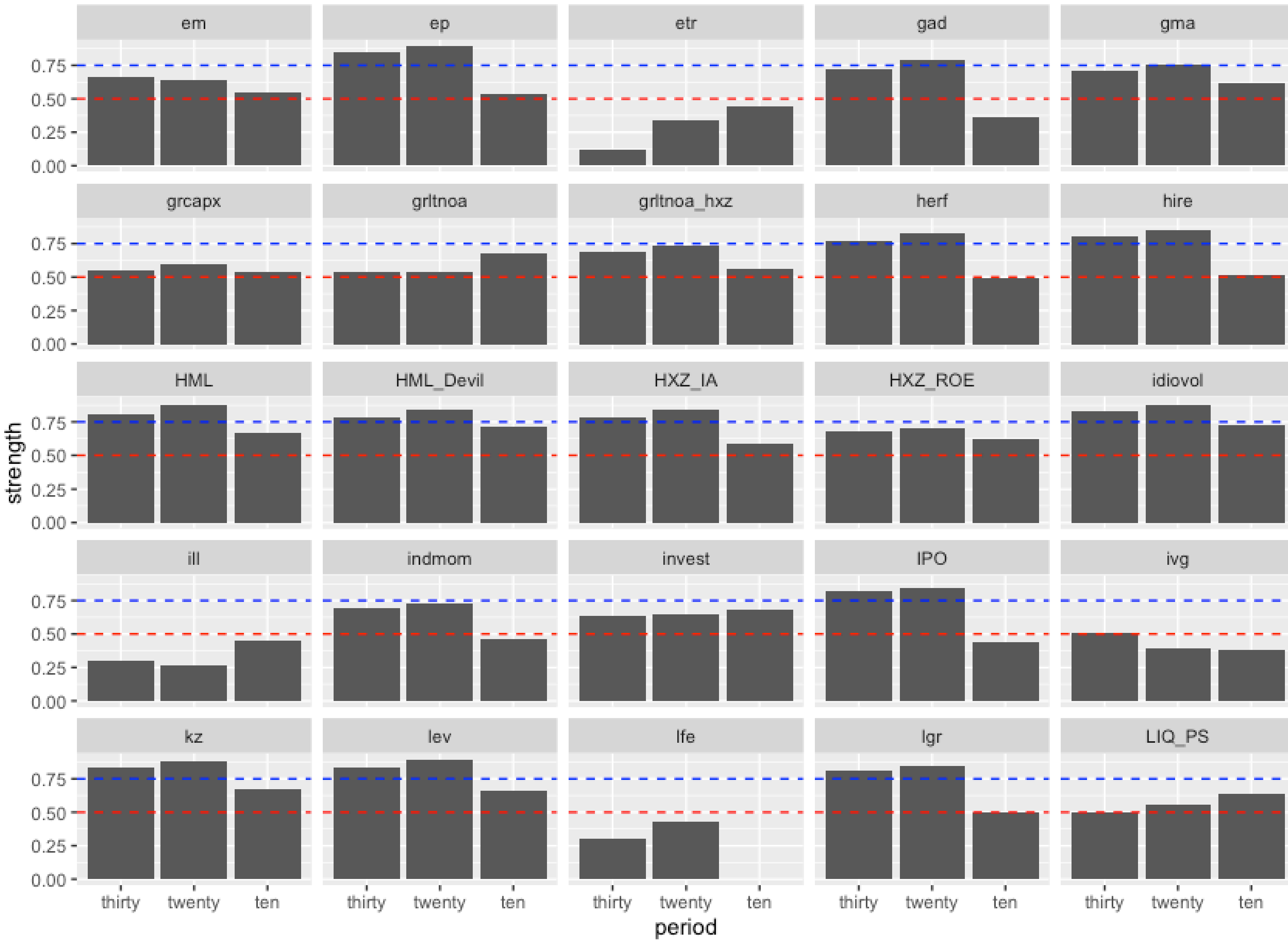
Notes: This table presents the estimated factor strength, using the decomposed thirty years data. The thirty year data set is decomposed into three subsets: January 1988 to December 1997, January 1998 to December 2007, and January 2008 to December 2017. For each data set, it contains 120 observations (t = 120), and 242 units (n = 242) The table also contains the full sample estimation results of factor strength, and the standard deviation among the three sub samples results. The table is ordered decreasingly base on the full sample factor strength.

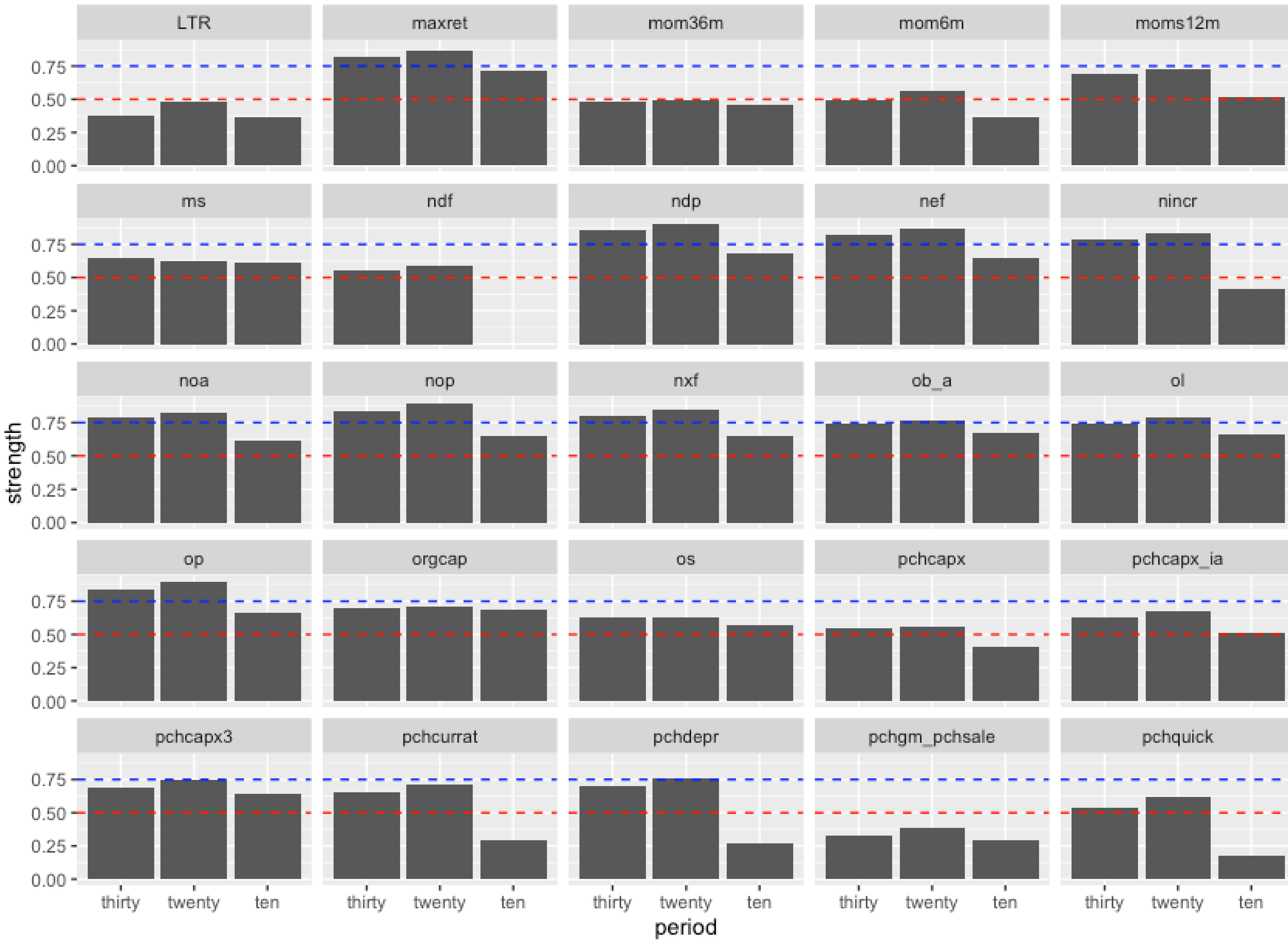
B.2 Strength Comparison Figures

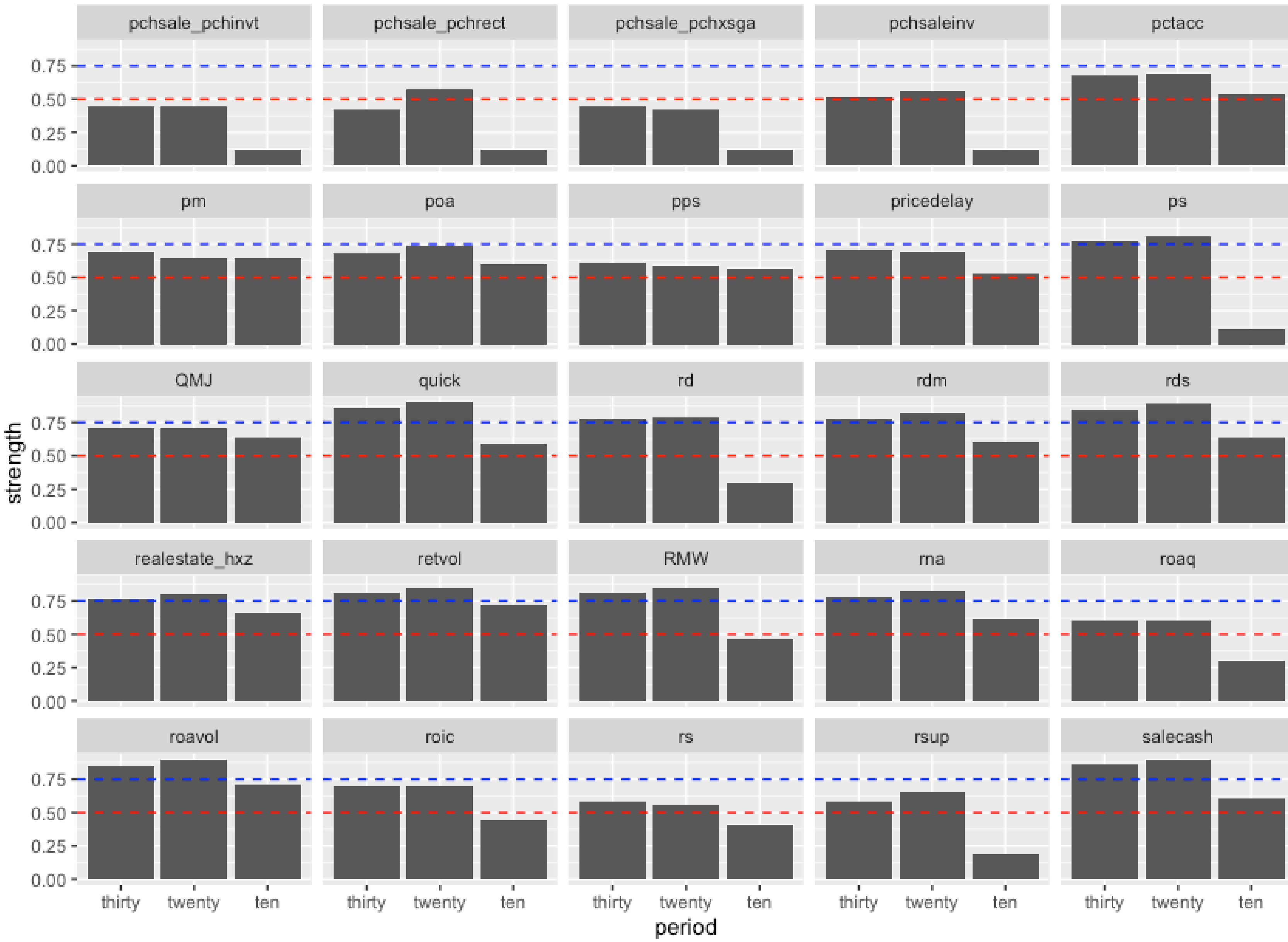
Figure 1: Strength Comparison

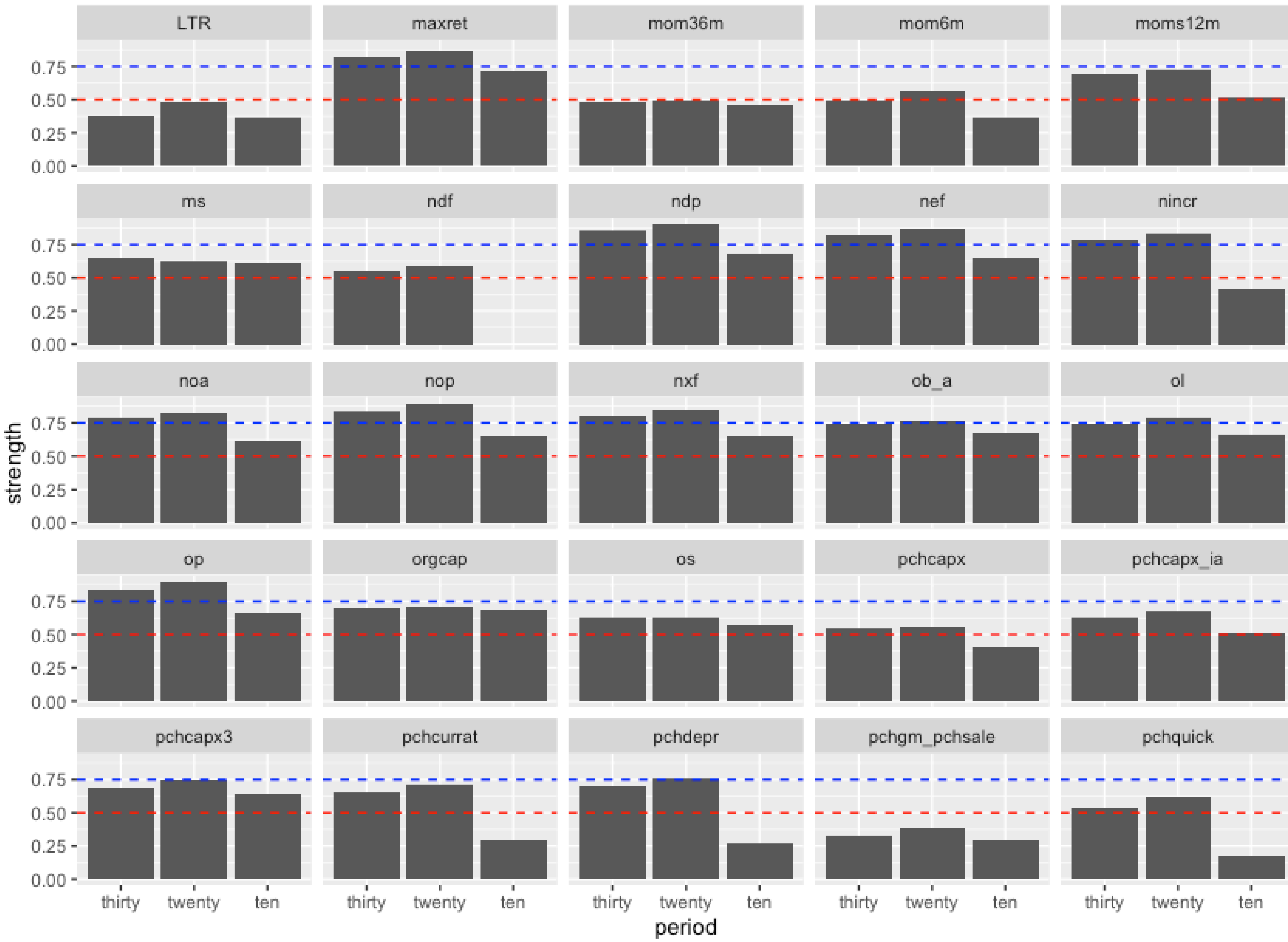



cinvest
cinvest_a
CMA
convind
cp
cto
currat
dcoa
dcol
depr
dfin
dfnl
dnca
dncl
dnco
dnoa
dolvol
dpia
dsti
dwc
dy
ear
ebp
egr
egr_hxz
strength
0.00
0.25
0.50
0.75
thirty
twenty
ten
period


em
ep
etr
gad
gma
grcapx
grltnoa
grltnoa_hxz
herf
hire
HML
HML_Devil
HXZ_IA
HXZ_ROE
idiovol
ill
indmom
invest
IPO
ivg
kz
lev
lfe
lgr
LIQ_PS
0.00
0.25
0.50
0.75
strength
thirty
twenty
ten
period

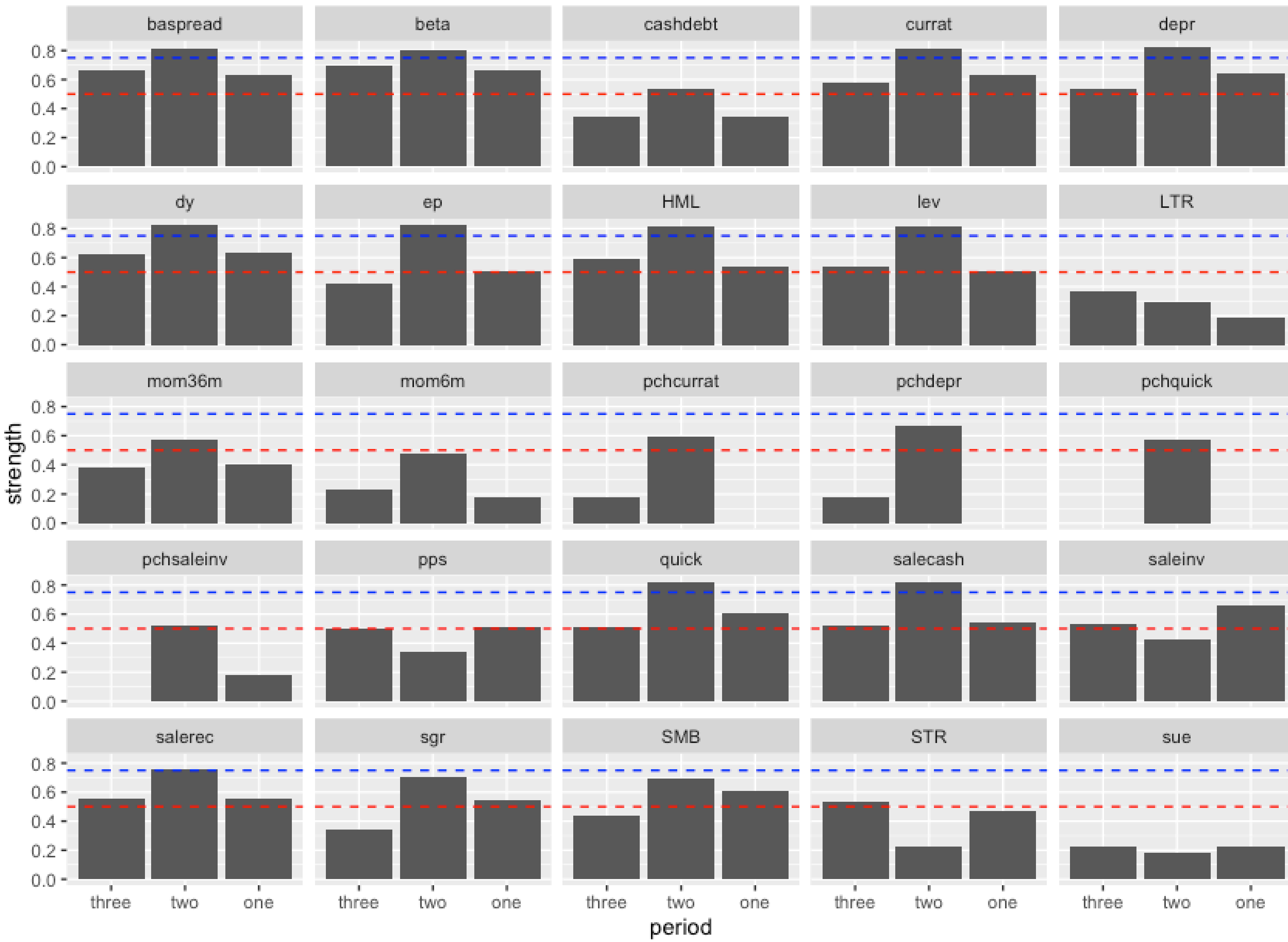

LTR
maxret
mom36m
mom6m
moms12m
ms
ndf
ndp
nef
nincr
noa
nop
nxf
ob_a
ol
op
orgcap
os
pchcapx
pchcapx_ia
pchcapx3
pchcurrat
pchdepr
pchgm_pchsale
pchquick
0.00
0.25
0.50
0.75
strength
thirty
twenty
ten
period

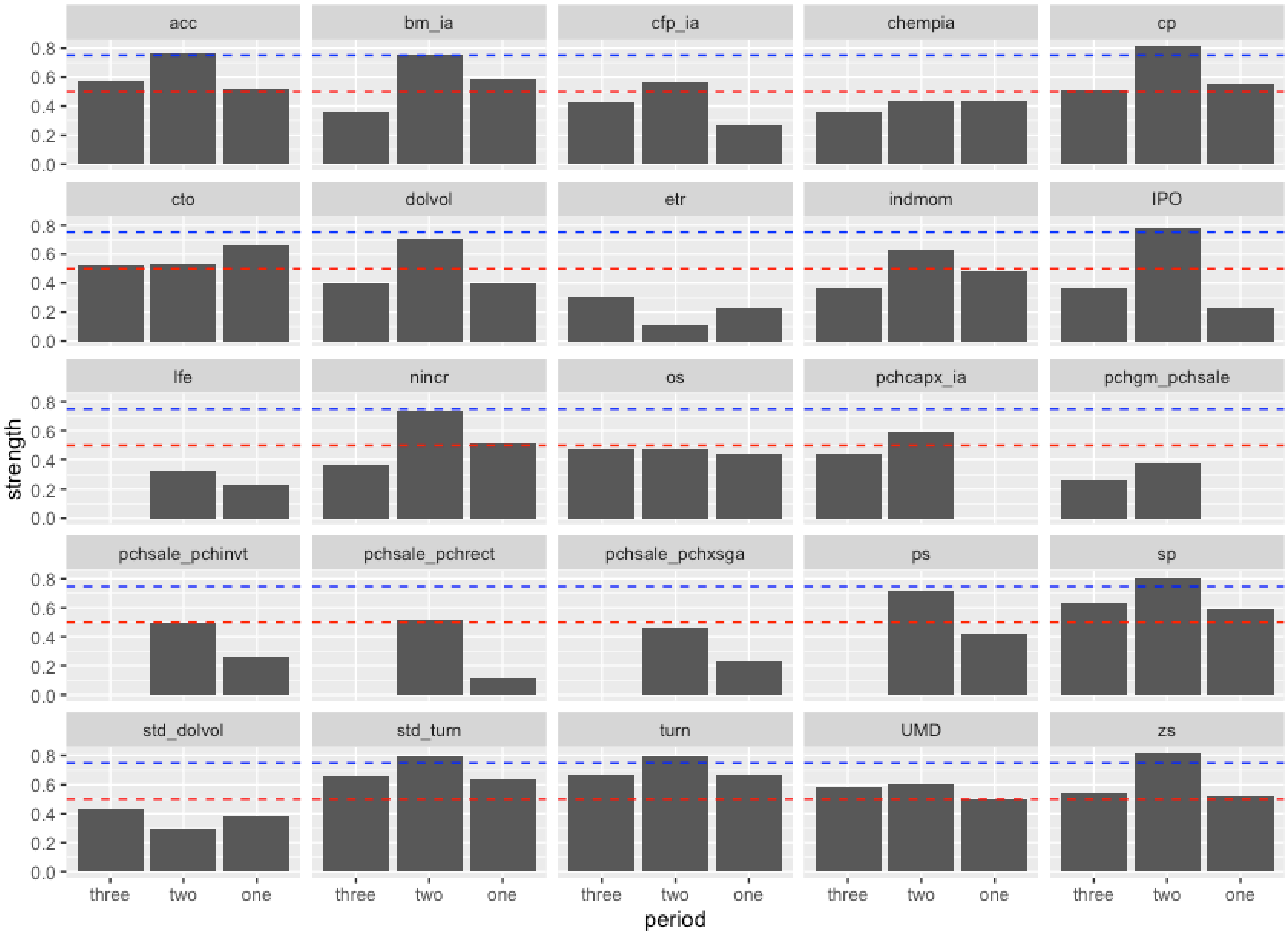

pchsale_pchinvt
pchsale_pchrect
pchsale_pchxsga
pchsaleinv
pctacc
pm
poa
pps
pricedelay
ps
QMJ
quick
rd
rdm
rds
realestate_hxz
retvol
RMW
rna
roaq
roavol
roic
rs
rsup
salecash
0.00
0.25
0.50
0.75
strength
thirty
twenty
ten
period

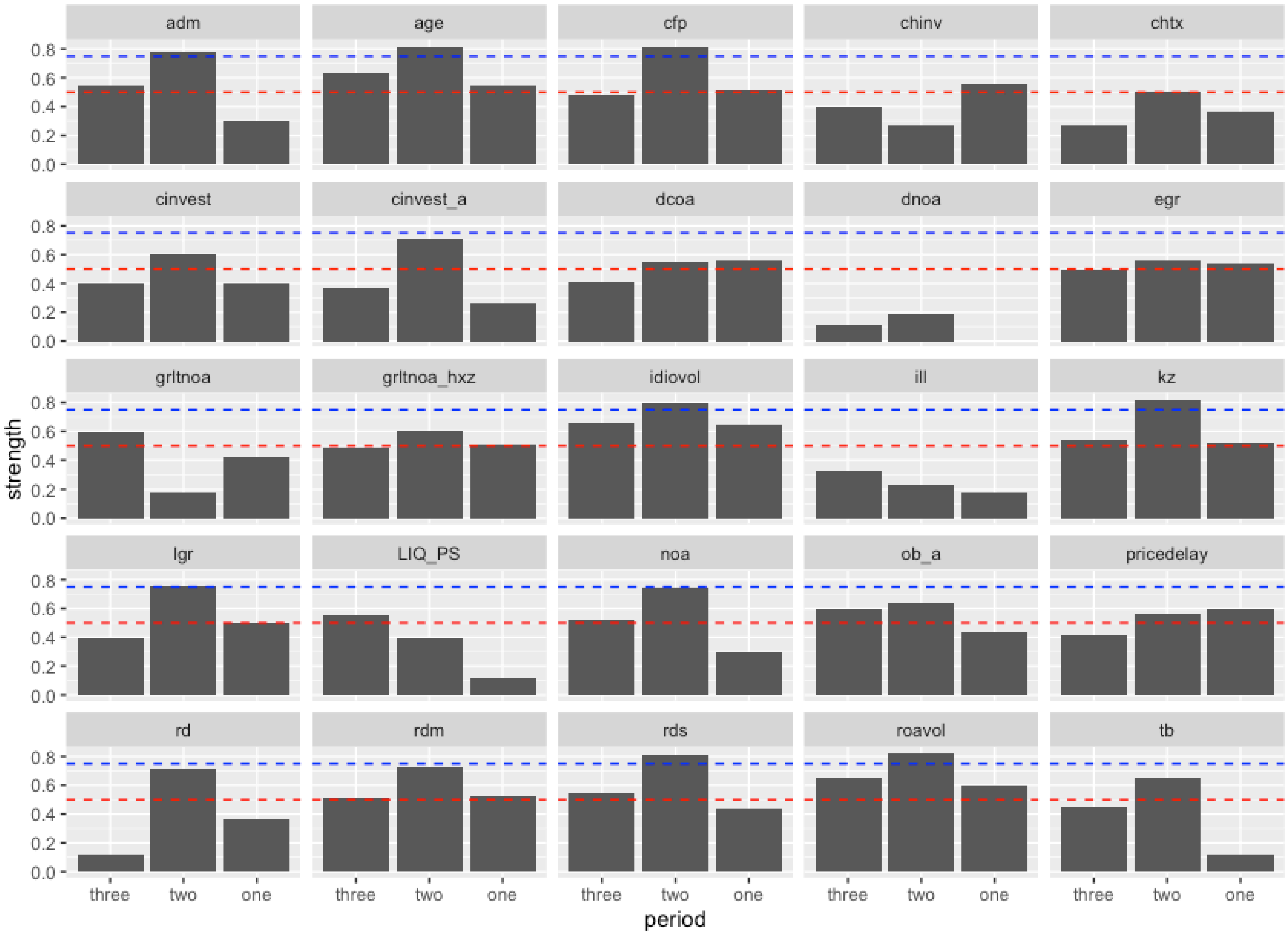


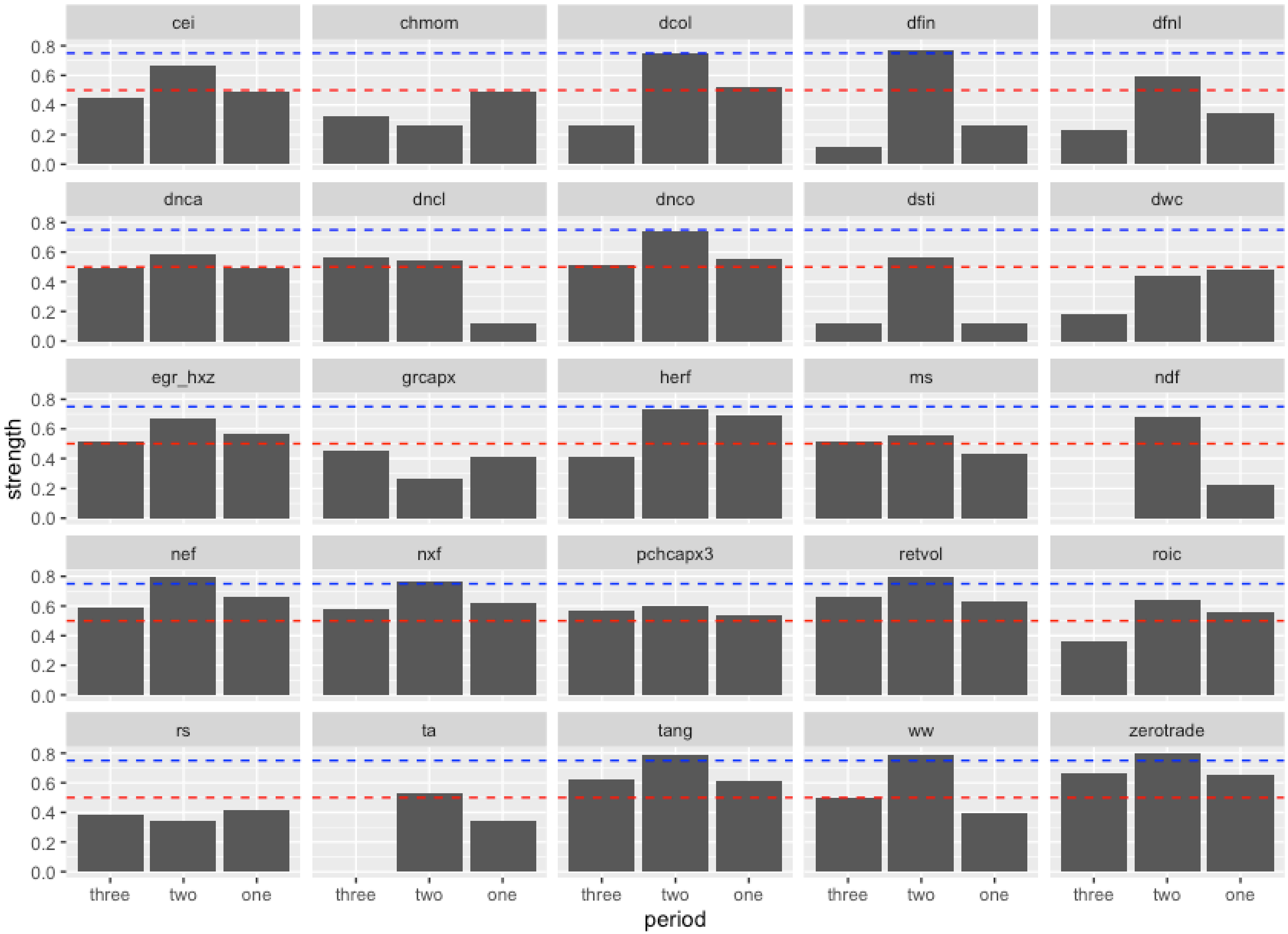
Notes: The figure compare the strength of every factor's strength in different data set. The x-axis indicates the data set: thirty is thirty years data set (January 1987 to December 2017), twenty is twenty year data set (January 1997 to December 2017), and ten is ten year data set (January 2007 to December 2017). The red dash line and blue dash line represent 0.5 and 0.75 threshold value respectively.

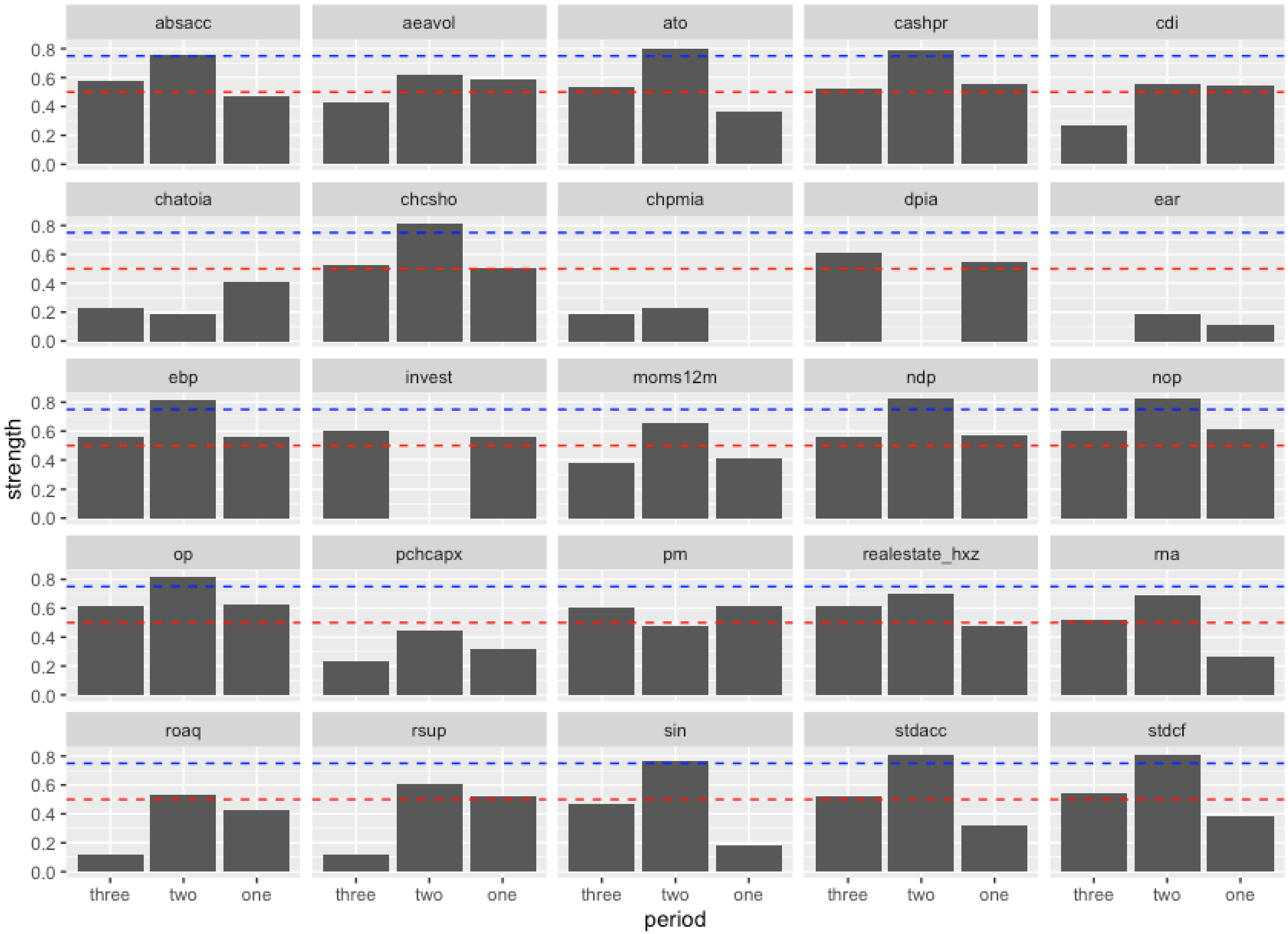
Figure 2: Thirty Year Decompose Comparison

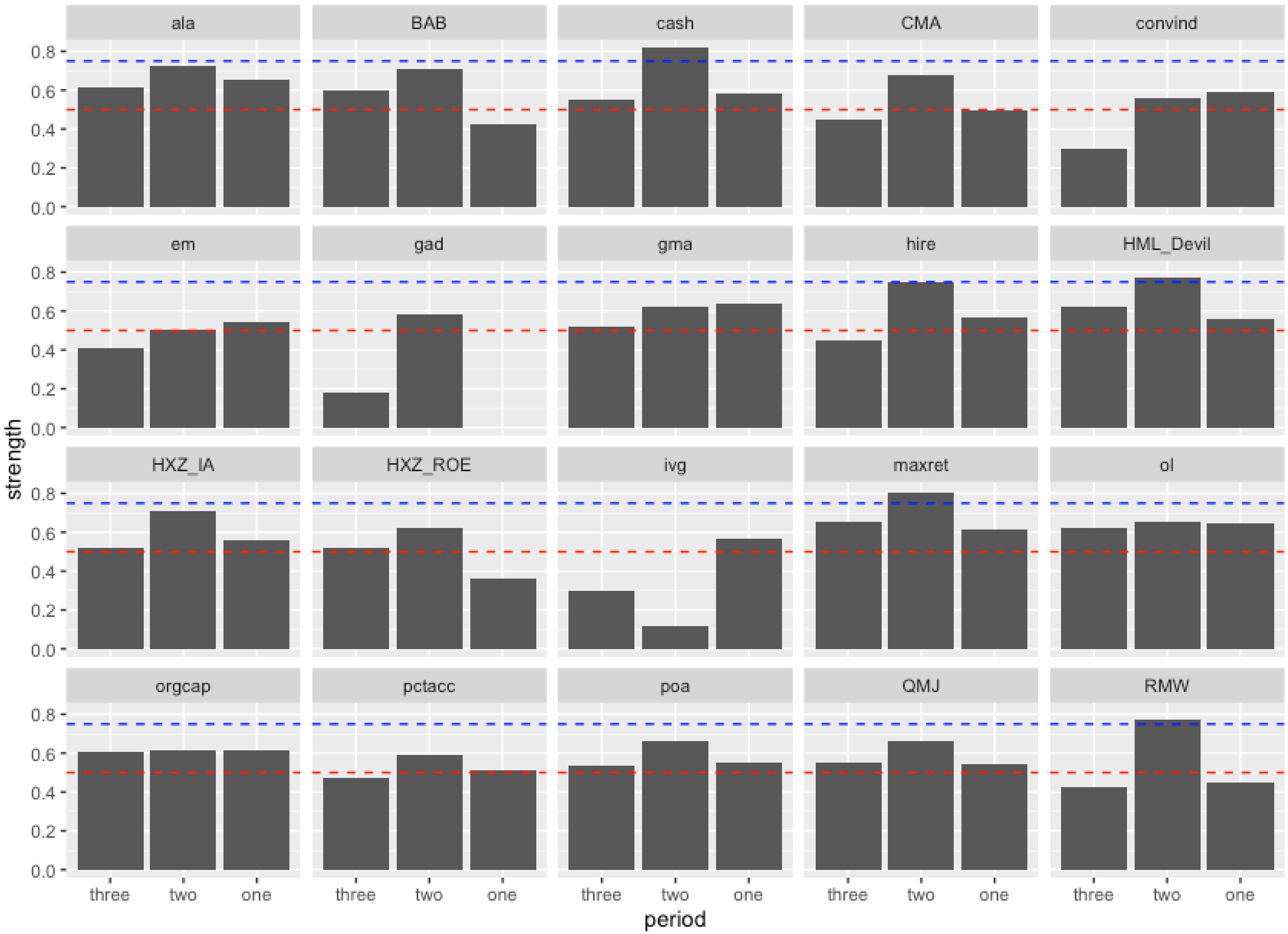



acc
bm_ia
cfp_ia
chempia
cp
cto
dolvol
etr
indmom
IPO
lfe
nincr
os
pchcapx_ia
pchgm_pchsale
pchsale_pchinvt
pchsale_pchrect
pchsale_pchxsga
ps
sp
std_dolvol
std_turn
turn
UMD
zs
0.0
0.2
0.4
0.6
0.8
strength
three
two
one
period


adm
age
cfp
chinv
chtx
cinvest
cinvest_a
dcoa
dnoa
egr
grltnoa
grltnoa_hxz
idiovol
ill
kz
lgr
LIQ_PS
noa
ob_a
pricedelay
rd
rdm
rds
roavol
tb
0.0
0.2
0.4
0.6
0.8
strength
three
two
one
period


cei
chmom
dcol
dfin
dfnl
dnca
dncl
dnco
dsti
dwc
egr_hxz
grcapx
herf
ms
ndf
nef
nxf
pchcapx3
retvol
roic
rs
ta
tang
ww
zerotrade
0.0
0.2
0.4
0.6
0.8
strength
three
two
one
period


absacc
aeavol
ato
cashpr
cdi
chatoia
chcsho
chpmia
dpia
ear
ebp
invest
moms12m
ndp
nop
op
pchcapx
pm
realestate_hxz
rna
roaq
rsup
sin
stdacc
stdcf
strength
0.0
0.2
0.4
0.6
0.8
three
two
one
period



Notes: The figure compare the strength of factor using subsample from the thirty year data.. The x-axis indicates the subsample data set: three is third decade (January 2007 to December 2017), two is second decade (January 1997 to December 2007), and one is the first decade (January 1987 to December 1997). The red dash line and blue dash line represent 0.5 and 0.75 threshold value respectively.