Factor Selection and Factor Strength

An Application to U.S. Stock Market Return

A thesis submitted for the degree of

Bachelor of Commerce (Honours)

Version: 1.2

by

Zhiyuan Jiang

with supervision from

Dr. Natalia Bailey

Dr. David Frazier

October 25, 2020

Abstract

Delete before submit: Remember what to include in the abstract: The identification of this paper, what this paper did, why we did so, what we found, and what are the meaning of those findings.

In this paper, we applied one new concept called factor strength to categorised the risk factors of CAPM model and applied a machine learning model called Elastic net to investigates the problem of how to select risk factors fro multi-factor CAPM model.

The high dimensions of factor groups increase the computational burden when applying the variable selecting algorithm, and a common feature of high-dimension factor group is that the correlation is pervasive therefore some selecting method is no longer appropriate in that scenario.

The factors strength helps to reduce the dimensionality of the candidates' factor group and provides a guideline to evaluates the performance of our factor selecting methods.

The Elastic net technique provides a potential solution for identifying factors and reducing redundant factors when constructing factor models. In brief, we find that the elastic net working effectively concerning reducing the redundancy in the factor groups.

This result proves the feasibility of applying Elastic net algorithm on selecting the risk factor for the CAPM model.

Keywords: CAPM, Factor Model, Factor Strength, Elastic Net, Lasso

Acknowledgement

I like to acknowledge...

Declaration

I declare that this research paper is my own work and has not been submitted in any form for another degree or diploma at any university or other institute of tertiary education. Information derived from the published and unpublished work of others has been acknowledged in the text and a list of references is given.

Zhiyuan Jiang

Contents

List of Figures

List of Tables

Chapter 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 the uncertainty of an asset's return into two parts: systematic risk part and asset specified idiosyncratic risk part. The systematic risk is captured by the market factor, which is represented by the difference between the average market return and the risk-free return of the market. Different risk factors contributed to price the idiosyncratic risk of different assets. Researchers (see Fama & French, 1992; Carhart, 1997; Kelly, Pruitt, & Su, 2019) have shown that by adding different risk factors into the model, CAPM can more precisely price the idiosyncratic risk. Because of this, identifying unique 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 different risk 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.

The high dimension factor group makes it hard for practitioners to find suitable factors when constructing the multi-factor CAPM model. And because of the dimensionality and computational burden, some traditional variable selection method like stepwise selection will not be applicable. When looking inside the factor group, we will find that some factors will only have a very weak or even no correlation with the asset's return.

This is because of the multiple-testing/data mining problem researchers will encounter when

they trying to identify new factors. The problem of multiple-testing will cause a false positive significant test results for factors, and mislead we to believe they can explain asset's risk. Hou, Xue, and Zhang (2018) argues that because researchers did not take the multiple testing into consideration when discovering new factors, a large portion of published factor results can not been replicated. Some evidences (see Kan & Zhang, 1999; Kleibergen & Zhan, 2015) have pointed out that, if the CAPM contains factor that has weak or even no correlation with the asset's return, the CAPM estimation will be distorted. Therefore, how to identify risk factors that can provide independent information about average return and risk become a crucial question in the field of finance and asset pricing (Cochrane, 2011).

In order to answer this problem, many scholars applied various methods to identify factors that can independently provide fresh information to explain risk-return relationship from a large existed factor pool. 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. In recent years, some other scholars are using machine learning methods to identify factors and eliminate redundant factors from a group of candidates. One stream of them has used a shrinkage and subset selection method called Lasso (Tibshirani, 1996) and the variations of it to find suitable factors. An example of such application is made by Rapach, Strauss, and Zhou (2013). They applied the Lasso regression, trying to find some characteristics from a large group, and then constructed a correspond CAPM model to predict the global stock market's return.

However, there is an additional challenge. Factors, especially in high-dimensions, are usually highly 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, because it can not distinct factors with strong correlation.

This leads to the main empirical question in this paper: how to select useful factors from a large group of possibly highly correlated candidates. We address this problem from two prospectives.

On the one hand, we employ a new idea called factor strength, trying to reduce the dimension of candidate factors by allocating them into smaller subgroups base on their estimated strength. The concept of factor strength is introduced by Pesaran and Smith (2019). They developed the idea to assess the significance of each factor. By measuring how many non-zero coefficient, or refer it as loadings in financial literature, a factor can generate for different assets, we can obtain the strength of every risk factor. With the factor strength, we can break down the high dimension group of candidate factors into small subgroups. It also provides us with a standard to evaluate the risk pricing performance of factor, enable us compare the risk-pricing ability across factors. So, we can use the strength as reference to evaluates how the factor selection method performed.

On the other hand, we use another variable selection method called Elastic Net (Zou & Hastie, 2005) to select factors from each subgroups. 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. Under the second approach, unlike Lasso, elastic net contains an extra penalty term, which enables it to avoid the problem of handling correlated features. This trait makes Elastic net fit for our purpose. We will assess and compare the methods in their selection of risk factors.

The rest of the thesis is organized as follows. In chapter 2, we go through some literatures relate with the CAPM model and methods about factor selection. Then in chapter 3, we will provide a detailed description of the concept of factor strength and the estimation method. In chapter 4, we set up a simple Monte Carlo simulation experiment to examine the finite sample properties of the factor strength estimator. Chapter 5 includes the empirical application regarding the factor strength, where we estimate the strength of each risk factors. We introduce and apply the Elastic net approach, alongside Lasso to select factors in chapter 6. Finally, we provides the conclusion and further discussion in chapter 7.

Chapter 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 Capital Asset Pricing Model (CAPM) builds up the connection between expected asset return and the risk. The original CAPM model only contains the market factor, which is denoted by the difference between average market return and risk-free return. ThenFama and French (1992) extend the model to contain size factor (SMB) and the value factor (HML). This three-factor models became popular in the finance industry. Carhart (1997), base on the Fama-French three factors model, added the momentum factor and makes it a new standard of factor pricing model. Some recent researches are attempting to extend the factor model even further. For instance, Fama and French (2015) create a five factors model base on their 1995 works by adding an investment factor and a profitability factor. They also created a six-factor model (Fama & French, 2018), by adding a momentum factor of their on version base on the five-factor models. Kelly et al. (2019) proposed a new method named Instrumented Principle Component Analysis (IPCA) which can identify latent factor structure. They applied the IPCA and constructed a six-factor model, claimed their six-factor model outperforms most of the sparse factor models, such as the five-factor models published by Fama and French in 2015.

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. The Fama-MacBeth two-stage regression (FM two-stage regression) introduced by Fama and MacBeth (1973) is a standard method when trying to estimates the CAPM and its multi-factor extension. Kan and Zhang (1999) found that the test-statistic of FM two-stage regression 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) showed how the addition of a spurious factor will distort the statistical inference of parameters, and misleads the researchers to believe that they correctly specified the factor structure, even when the degree of misspecification is arbitrarily large. 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. To address the problem of misspecified factor-return relationship, Gospodinov, Kan, and Robotti (2014) proposed a factor selection procedure, which bases on their statement, can eliminates the falsely presented factor robustly, and restores the inference.

Finally, of interest in this thesis is the large dimension of potential factors. Harvey and Liu (2019) documented over 500 published risk factors, and they indicated that more factors are discovered every year. Among all those risk factors, Hou et al. (2018) tried to replicates 452 of them, and they find only 18% to 35% factors are reproducible. For the reasons and findings above, this thesis also 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. Some other methods are also developed to selecting factors. Barillas and Shanken (2018) introduces a Bayes test procedures. It enables researchers to compare the probabilities of a collection of potential models, which can be constructed after giving a group of factors. In order to identify factors risk-pricing ability, 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 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 results. In recent years, machine learning algorithms have become popular in the finance studies, and various methods are adopted when selecting factors for the factor model. Lettau and Pelger (2020) apply Principle Components Analysis (PCA) when investigating the latent factor of the model. Lasso, been innovated by Tibshirani (1996), is a popular algorithm which can eliminate redundant features. The derivation of Lasso has become increasingly popular in the factor selection. For example, Feng, Giglio, and Xiu (2019) used the double-selected Lasso method (Belloni, Chernozhukov, & Hansen, 2014), and Freyberger, Neuhierl, and Weber (2020) used a grouped lasso method (Huang, Horowitz, & Wei, 2010) 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.

Chapter 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:

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

Here, the $\boldsymbol{M}_\tau = \boldsymbol{I}_T - T^{-1}\boldsymbol{\tau}\boldsymbol{\tau}'$, and the $\boldsymbol{\tau}$ is a $T \times 1$ vector with every elements equals to 1. $\boldsymbol{f}$ and $\boldsymbol{x_i}$ are two vectors with: $\boldsymbol{f} = (f_1, f_2 \cdots, f_T)'$ $\boldsymbol{x_i} = (x_{i1}, x_{i2}, \cdots, x_{iT})$. 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:

$$x_{it} = a_i + \sum_{j=1}^{k} \beta_{ij} f_{jt} + v_{it} = a_i + \boldsymbol{\beta}_i' \boldsymbol{f}_t + v_{it} \tag{7}$$

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 $\boldsymbol{\beta}_i = (\beta_{i1}, \beta_{i2}, \cdots, \beta_{ij})'$ and $\boldsymbol{f}_t = (f_{1t}, f_{2t} \cdots, f_{jt})$. We employed the same strategy as above, after running OLS and obtain the:

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

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:

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

Here, $\boldsymbol{f}_{jt} = \left(f_{j1}, f_{j2}, \ldots, f_{jT}\right)', \boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iT})', \boldsymbol{M}_{F_{-j}} = \boldsymbol{I} - \boldsymbol{F}_{-j}\left(\boldsymbol{F}_{-j}' \boldsymbol{F}_{-j}\right)^{-1} \boldsymbol{F}_{-j}'$, and $\boldsymbol{F}_{-j} = \left(\boldsymbol{f}_{1t}, \ldots, \boldsymbol{f}_{j-1t}, \boldsymbol{f}_{j+1t}, \ldots, \boldsymbol{f}_{mt}\right)'$ 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.

Chapter 4

Monte Carlo Simulation

In this chapter, we set up several simple Monte Carlo simulation experiment to study the finite sample properties of factor strength $\hat{\alpha}_j$. Through the simulation, we compare the property of the factor strength in different settings.

4.1 Simulation Design

The experiments is designed 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. The market factor is defined as the difference between the average market return, and the risk free return. Here r_{mt} is the average market return of hypothetically all assets in the universe. Now consider the following data generating process (DGP):

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

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.

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

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

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

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

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 chapter (3.2). This step reflects the fact that only $[n^{\alpha_1}]$ or $[n^{\alpha_2}]$ 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 four different experiments set up:

Experiment 1 (single factor, normal error, no correlation) *Set β_{i2} from (8) 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.*

Experiment 4 (two factors, normal error, strong 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 0.7. The factor strength for the first factor $\alpha_1 = 1$ all the time, and α_2 varies.*

The factor strength in experiment one is estimated using the method discussed in chapter (3.2), and for the rest of experiments, we use the method from chapter 3.3. 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 $\hat{\alpha}$.

$$bias = \frac{1}{R}\sum_{r=1}^{R}(\alpha - \hat{\alpha}_r)$$

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).

$$z_{\hat{\alpha}_j:\alpha_j} = \frac{(\ln n)\left(\hat{\alpha}_j - \alpha_j\right) - p\left(n - n^{\hat{\alpha}_j}\right)n^{-\delta-\hat{\alpha}_j}}{\left[p\left(n - n^{\hat{\alpha}_j}\right)n^{-\delta-2\hat{\alpha}_j}\left(1 - \frac{p}{n^{\delta}}\right)\right]^{1/2}} \quad j = 1,2 \tag{9}$$

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:

$$size = \frac{\sum_{r=1}^{R} \mathbf{1}(|z_{\hat{\alpha}_j:\alpha_j}| > 1.96|H_0)}{R} \quad j = 1,2, \tag{10}$$

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 Findings

We report the results in Table (A.1), (A.2), (A.3), and (A.4) in Appendix A.

Table (A.1) 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 (9), 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 no correlation setting, weak correlation setting, and the strong correlation setting. The result of no correlation settings is shown in the table (A.2), table (A.3) shows the result when the correlation between two factors is 0.3, and the table (A.4) presents the result of 0.7 correlation setting. Same as the single factor scenario, 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 (A.3)). 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. However, it is worth notice that in the strong correlation setting, the size of the test is extremely big when the time spam is relatively short, and the test size will not improve even we increase the sample size. Once we increase the t, the size of the test will reduce dramatically, and when we have the thirty year time period, the size of the test are almost all below the 0.05 threshold.

Chapter 5

Empirical Application: Factor Strength

Researchers and practitioners have been using the CAPM model (Sharpe, 1964; Lintner, 1965; 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 chapter, we will use the method introduced in the chapter 3 to estimates the factor strength of 146 candidate factors.

First, we introduce the data set used in this empirical chapter, and the setting for the estimation process. Then, we discuss the findings from the estimation results. The estimated factors strength provides us a starting point when applied the Elastic net method in the chapter 6.

5.1 Description of Data for Factor Strength Estimation

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 com-

[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/.

ponents 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 (5.1) below.

Table 5.1: 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 stocks in the U.S. market, incorporated by CSRP. Each factor contains observations from January 1988 to December 2017.

[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.

5.2 Setting for Factor Strength Estimation

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

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

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

Here r_{it} is the return of asset i at time t, and rf_t is the risk free return. Therefore, $r_{it} - rf_t$ is the excess return of the asset i. $r_{mt} - rf_t$ 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.

5.3 Factor Strength Estimation and Discussion

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 5.1, and rank those strength from strong to weak, alongside the market factor strength, in the table (B.1).

We first look at the market factor strength under the single-factor CAPM setting. (see table 5.2)

Table 5.2: 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	0.991	0.996

As we expected, the estimated strength of market factor under all three scenarios shows consistently strong results. All three market factor 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 of the market than the short run. Then, we turn to the double factor CAPM setting. Under the double factor setting, we estimates every companies' stock return using market facto plus another risk factors. Therefore we will obtain the market factor strength for each single risk factors. So, we calculate the average market factor here:

$$\bar{\alpha}_m = \frac{1}{145}\sum_{i=1}^{145}\alpha_{im}$$

α_{im} is the estimated market factor strength for the i^{th} risk factor setting. Table5.2 shows that the estimated strength is consistent with the single factor settings. All three data sets provides extremely strong results, strengths are close to one. Overall, the market factor results indicates that the market factor can generate significant loadings for almost every companies' return at any time. Such conclusion fits with the financial theory states that individual stocks will in general move with the market. 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 0.937. In the thirty-year scenario, the salecash (sales to cash) is the strongest with strength 0.948.

Table 5.3: Proportion of factor within certain strength range

Strength Level	10 Year Data Proportion	20 Year Data Proportion	30 Year Data Proportion
[0.9, 1]	0%	21.4%	23.4%
[0.85, 0.9)	0%	17.9%	17.9%
[0.8, 0.85)	0%	7.59%	6.21%
[0.75, 0.8)	0%	11.7%	17.9%
[0.7, 0.75)	7.59%	8.28%	7.59%
[0.65, 07)	15.9%	8.28%	2.76%
[0.6, 0.65)	17.9%	5.52%	8.97%
[0.55, 0.6)	13.1%	6.21%	2.76%
[0.5, 0.55)	8.97%	4.83%	4.14%
[0, 0.5)	36.6%	8.28%	8.28%

When comparing the proportion of factors with strengths falling in different intervals between 0 and 1 (see table (5.3)), we can find that when using 0.8 as a threshold, there are over forty-five per cent factors in the twenty-year and thirty-year result exceeds this threshold. 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 four times higher than the twenty and thirty years proportion.

When look at the ranking, we found that there are three factors: ndq (Net debt-to-price), salecash (sales to cash), and quick (quick ratio) who are ranking top three in both twenty year data set result and thirty year data set result. We would expect when applying the elastic net method with the twenty-year and thirty-year data set, those three factors with the market factors would be selected.

Another interesting finding is that the roavol (Earnings volatility, 10th of ten-year result, 7th of twenty-year result, 5th of thirty-year result), age (Years since first Compustat coverage, 11th of ten-year result, 9th 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.

Table 5.4: Selected Risk Factor with Strength: top 15 factors from each data set and three well known factors.

Ten Year			Twenty Yera			Thirty Year		
Rank	Factor	Strength	Rank	Factor	Strength	Rank	Factor	Strength
1	beta	0.749	1	ndp	0.937	1	salecash	0.948
2	baspread	0.730	2	quick	0.934	2	ndp	0.941
3	turn	0.728	3	salecash	0.933	3	quick	0.940
4	zerotrade	0.725	4	lev	0.931	4	age	0.940
5	idiovol	0.723	5	cash	0.931	5	roavol	0.938
6	retvol	0.721	6	dy	0.929	6	ep	0.937
7	std_turn	0.719	7	roavol	0.929	7	depr	0.935
8	HML_Devil	0.719	8	zs	0.927	8	cash	0.934
9	maret	0.715	9	age	0.927	9	rds	0.931
10	roavol	0.713	10	cp	0.926	10	dy	0.927
11	age	0.703	11	ebp	0.926	11	currat	0.927
12	sp	0.699	12	op	0.925	12	chcsho	0.927
13	ala	0.699	13	cfp	0.924	13	lev	0.926
14	ndp	0.686	14	nop	0.924	14	stdacc	0.926
15	orgcap	0.686	15	ep	0.923	15	cfp	0.925
20	UMD	0.678	29	HML	0.905	39	HML	0.894
24	HML	0.672	76	SMB	0.770	68	SMB	0.804
87	SMB	0.512	89	UMD	0.733	96	UMD	0.745

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 closely around 0.9, none of the other factors in any data set shows strength higher than 0.85.

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 0.905, for SMB and UMD the strength are 0.770 and 0.733 respectively. Comparing with the twenty-year results, the thirty-year estimated strength change slightly, HML decreases to 0.894, SMB is 0.804 and UMD has strength 0.745. Therefore, when using the strength as a criterion, we may only select the value factor to incorporate in the CAPM model when having twenty and thirty-year data.

In general, we found that the twenty-year and thirty-year data sets provides similar esti-

mated strength, while, in contrast, the estimated strengths from ten-year data set are significantly weaker. Therefore, 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 (B.2) and figure (B.2).

In general, we can conclude that about 80% 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 (B.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 dwc (change in net non-cash working capital) factor (see figure B.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.

Chapter 6

Empirical Application: Elastic Net

As we introduced before, the surging number of risk factors post question of which factors can independently provides unique information to explain the risk-return relationship. (Cochrane, 2011). In this chapter, we employed the Elastic net algorithm to discuss this problem. The application is build on the basis of factor strength we estimated in the previous chapter. Factor strength provides a criterion to reduce the dimensions of potential factor groups. The factor strength also works as a reference when evaluating the performances of algorithms. We would expect the method tend to select factor with strong strength than weak one. For the rest of this chapter, we first briefly introduce the core idea of the Elastic net method, explained what's different between Elastic net and other factor selection methods, especially the Ridge regression and Lasso regression. Then, we provide a full discussion of how to select the tuning parameter with regard to the Elastic net method when using r package *glmnet* to imply the Elastic net. Finally, we compare and discuss the differences selection results of both elastic net methods and Lasso.

6.1 Brief Introduction to Elastic Net

Elastic net,introduced by Zou and Hastie (2005), is a penalised linear regression method which developed on the basis of the Lasso regression (Tibshirani, 1996) and Ridge regression. To illustrate the application of the Elastic net method in our research, recall the multi-factor model

(7) we discussed in chapter 3.3. When applying the OLS to estimate the factor loading $\boldsymbol{\beta}_i$ of model (7), we targeting minimise the Residual Sum of Squares (RSS):

$$\hat{\boldsymbol{\beta}}_i = \arg\min\{(x_{it} - \hat{a}_{iT} - \hat{\boldsymbol{\beta}}_i' \boldsymbol{f}_t)^2\}$$

The OLS method will consider all factors proposed by users when constructing the multi-factor CAPM model. This means that even factor with weak strength, which can not bring any new information to price the risk of assets will generate loadings under the OLS method. The Elastic net method, although also focusing on minimising RSS, including two extra penalty terms inside the loss function.

$$\hat{\boldsymbol{\beta}}_i = \arg\min_{\beta_{ij}}\{(x_{it} - \hat{a}_{iT} - \hat{\boldsymbol{\beta}}_i' \boldsymbol{f}_t)^2 + \lambda_2 \sum_{j=1}^{k} \hat{\beta}_{ij}^2 + \lambda_1 \sum_{j=1}^{k} |\hat{\beta}_{ij}|\} \tag{11}$$

Here, the estimated $\hat{\boldsymbol{\beta}}_i$ loading value is subject to two penalty terms: the L^1 norm $\lambda_1 \sum_{j=1}^{k} |\hat{\beta}_{ij}|$ and the L^2 norm: $\lambda_2 \sum_{j=1}^{k} \hat{\beta}_{ij}^2$. The Elastic net estimation, in essence, is a combination method of Ridge regression and the Lasso regression. If setting $\lambda_1 = 0$, the Elastic net will collapse into the Lasso regression, and if $\lambda_2 = 0$, we will obtain result of the Ridge regression. In the empirical application of this paper, the Elastic net estimation uses the following form (Friedman, Hastie, & Tibshirani, 2010):

$$\hat{\boldsymbol{\beta}}_i = \arg\min_{\beta_{ij}}\{\frac{1}{2N}(x_{it} - \hat{a}_{iT} - \hat{\boldsymbol{\beta}}_i' \boldsymbol{f}_t^2) + \phi P_\theta(\boldsymbol{\beta}_i)\} \tag{12}$$

$$P_\theta(\boldsymbol{\beta}_i) = \sum_{j=1}^{k} [(1-\theta)\beta_{ij}^2 + \theta|\beta_{ij}|] \tag{13}$$

Here we call the $P_\theta(\cdot)$ as the elastic net penalty. θ acts as the turning parameter to determine how will the elastic net penalty is combined by the L^1 and L^2 norms. When set $\theta = 1$, we have $P_\theta(\boldsymbol{\beta}_i) = \sum_{j=1}^{k} |\beta_{ij}|$ which is identical to the L^2 norm, therefore we have the Elastic net collapse to the Lasso regression. Similarly, when setting $\theta = 0$, we have the Elastic net collapse to the ridge regression, and when $\theta = 0.5$, the Elastic net is behave like the combination of ridge and

Lasso. The other tuning parameter ϕ decides how strong the penalty terms is. If $\phi = 0$ the elastic net will become the OLS estimation.

In this study, we use the data introduced in chapter 5.1, and the estimated factor strength from the chapter 5. We will briefly review risk factors with their strength in the next chapter. More specifically, we allocates the 145 risk factors into six subgroups base on there thirty-year estimated strength. For each subgroup, we want to investigate how will the Elastic net algorithm, alongside the Lasso regression, select the risk factors for company stock. We would expect that when facing factor with strong strength, the algorithm will construct a dense factor model, and with the factor strength decrease, the density will decrease simultaneously. Notice that in order to simplify the application of the Elastic net, instead of using the excess return of stock directly, we first run an OLS regression between the market factor and the excess return of each company. And then, we collect the residuals of the OLS estimation, use the residuals as the left hand side term of the multi-factor CAPM model. This step helps us to remove the potential influence of market factors, allowing the algorithms only focussing on the risk factors.

Now, the main challenge of applying the elastic net algorithm is to select the appropriate tuning parameters θ and ϕ, and we will discuss the choices of tuning parameter in the following section.

6.2 Properties of Risk Factors

In this chapter, we briefly review the properties of the risk factors series. The risk factors data set contains 145 risk factors at the monthly frequency for the period from July 1976 to December 2007.[1] As a standard practice of time series data, we exam the stationarity of the risk factor series by employing Augmented Dick-fuller (ADF) test (Dickey & Fuller, 1979), Phillips-Perron (PP) test (Phillips & Perron, 1988), and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test (Kwiatkowski, Phillips, Schmidt, & Shin, 1992). Overall, the ADF test and PP test provides the same conclusions that all 145 risk factor series does not contain unit root. The KPSS test,

[1] For how those factors are constructed, please view Feng et al. (2020)

however, disagrees with the ADF and PP test. If we take 0.05 as the threshold of p-value, the KPSS test concludes that there are six factors that contain unit root in their process. We also present the correlation heat map of all 145 risk factors in the Figure B.3 at the appendix B.3. All risk factors are sorted base on their estimated factor strength. The dark lower-left area of the Figure B.3 indicates that factor with strong estimated strength presents a high correlation with other strong factors. With the factor strength decrease, the correlation coefficient among factors decreases significantly. If focusing on the upper-right corner of the figure, where weak factors are clustering, we can see that the correlation coefficients are close to zero. By observing the upper-left corner and the lower-right corer, we can observe that the correlation between weak factors and strong factors are also very low

6.3 Tuning Parameter

In this empirical application, we use the R package *glmnet* (Friedman et al., 2010; Simon, Friedman, Hastie, & Tibshirani, 2011).

From the equation (12) and (13) we know that the estimation of Elastic net dependent on two tuning parameters θ and ϕ. The *glmnet* package provides function to select the ϕ automatically. This selection is base on the minimisation of Mean Squared Error (MSE), using cross-validation. However, the package does not provide aid on determining which value of θ parameter is optimal. Therefore, we adopt the following strategy to select our tuning parameters ϕ and θ:

1. Prepare a sequence of θ values, from 0: ridge regression, to 1: Lasso regression with the step of 0.01

2. Randomly assign 90% of the data set as the training set and the rest 10% as the test set.

3. For each of the θ value, we fit the corresponding Elastic net model using the training set, with function picked ϕ values.

4. Base on the θ and ϕ values select, we produce the predicted values using the test data, and calculate the MSE between the true values and predicted values.

5. We select the $\theta - \phi$ combination which minimises the MSE.

We repeat the above procedures 2000 times for each factor strength group and calculates the $\bar{\theta}$ by averaging every θ we obtained from the repetition. Due to the computational burden, when implying step 2, instead of using the full sample of stocks and factors, we randomly select 50 companies from 242 all companies, and 10 factors from each subgroups. The selected tuning parameter θ values are shown in table 6.1.

Table 6.1: Estimated Optimal θ values for different factor groups

Factor Group	(0, 0.5]	(0.5, 0.6]	(0.6, 0.7]	(0.7, 0.8]
Selected θ value	0.377	0.401	0.429	0.411
Factor Group	(0.8, 0.9]	(0.9, 1]	Mix	Random
Selected θ value	0.396	0.413	0.448	0.431

In order to fully investigate the behaviours of parameter tuning, on the basis of the six sub-groups, we create tow addition groups. The Mix group contain five highly strong factors: factor with strength higher than 0.9, and five weak factors: factor with strength lower than 0.5. The Random group consist of ten randomly selected factors from all 145 risk factors. From table 6.1 we can see that the selected θ value in general increase with the factor strength increase. For weak factor group, the selected parameter value is 0.377, close to the ridge regression. While for the strong factor group, the value is 0.413. The mix factor group has the highest θ value of 0.448.

Such a pattern of the θ value, however, does not follow what we expected. The definition of factor strength indicates that factor with strong strength is able to produce more significant loadings, in other words, can explain more assets' risk-return relationship. Therefore, when using the above procedure to decide tuning parameter, we would expect that for factor groups with lower strength, like groups with strength smaller than 0.5, the selected θ parameter will be

close to one, or larger than other groups' selected θ. This is because factor with weak strength will only provide limited pricing power, and therefore may be recognised by the algorithm as redundant variables. When θ is closer to unity, the elastic net is behaved more like a Lasso, which will eliminate variables provides limited explaining power. In contrast, when the group has stronger strength, the θ will approach closer to zero, leads the Elastic nets more like Ridge, which will not eliminate any variables, but only reduce the coefficient. So we would expect the θ value to increase with the group strength decrease.

We prepared several potential explanations for this result. First, the MSE is not an ideal criterion for selecting the tuning parameter. The MSE for all $\theta - \phi$ combinations are very close to each other. All MSE results show similar values around 64. Second, because of the estimation method we used, the market risk has already been absorbed by the market factors. Then for the strong factors, we would expect that for any single of them, those strong factor can individually explain most of the idiosyncratic risk. Therefore, when we ask any ten strong factors to determine the risk simultaneously, it is possible that very few of the ten factors can explain most of the risk and therefore the other risk factors will be recognised by the algorithm as redundant. But if all factors are weak, it is possible that there exist some linear combinations among weak factors provides enough explaining power for the idiosyncratic risk. Therefore, those weak factors will be reserved by the algorithm, and hence, the parameter θ will close to zero, indicates a more Ridge like regression.

6.4 Elastic Net Findings

We applied the elastic net algorithm with tuning parameters estimated in the previous chapter using the thirty-year data set. The reasons why we use the thirty-year data instead of ten or twenty years is because the Monte Carlo simulation result (see chapter 4.3 and Appendix A) indicates us that the estimation accuracy of factor strength will improve significantly with a decreasing of size of the test, as the increasing of time spam. Even for the data set with limited observations, estimation shows high accuracy when t is large. Therefore, we use the factor strength estimated

from the thirty-year data set.

We divided the 145 risk factors into six groups base on their factor strength. We also randomly selected ten factors from weak factor group (factor with strength less than 0.5), and ten factors from strong factor group (factor with strength above 0.9) to form a mixed factor group. For each factor groups, we run two regression: the elastic net regression with θ values presented in table 6.1, and the Lasso regression with $\theta = 1$. Instead of running a pooled regression, we run the elastic net and Lasso for each individual company, and record the result of factor selection of every stock. First, we focusing on the general behaviours of factor selection between our two methods. Table 6.2 presents the average factor selection amount for each factor groups of two selection methods. We can see that the factor model selected by the Lasso regression,

Table 6.2: Average factor selection proportions and factor selection counts of Elastic Net and Lasso

Factor Group	(0,0.5]	(0.5, 0.6]	(0.6, 0.7]	(0.7, 0.8]	(0.8,0.9]	(0.9,1]	Mix
Factor Amount	12	10	17	37	35	34	20
Avg EN selection amount	2.11	4.47	8.67	14.67	13.51	12.37	8.45
Avg EN selection proportion	17.5%	44.73%	51.00%	39.65%	38.61%	36.38%	42.28%
Avg Lasso selection amount	2.06	3.87	8.43	13	12.19	10.46	7.26
Avg Lasso selection proportion	17.2%	38.76%	49.60%	35.14%	34.83%	30.75%	36.27%

on average, is more parsimonious than the model selected by the Elastic net. And this trend becomes more and more obvious with the factor strength increase. For the weak factor group with strength less than 0.5, Elastic net and Lasso provides a similar answer: only two factors are selected out of twelve candidates. While when focusing factor with strength above 0.9, Lasso will select almost 2 fewer factors than the Elastic net. Such outcome is not surprising, since the tuning parameter θ of Elastic net is closer to 0 than 1, which means that the Elastic net algorithm in our application tend to keep the factors even though they can provide very limited information. Unexpectedly, the proportion of factor selection to the strong-factor group is significantly lower than the semi-strong group. For the factor group with strength between 0.6 and 0.7, both Elastic net and Lasso select almost half of 17 candidates factors. But when facing strong factors with strength above 0.9, Elastic net on average only keep about 36% factors.

We then compare every single stock's factor selecting decision made by Lasso and Elastic

net, and calculates in what degree those two methods will agree with each other. For every single company, if both Elastic net and Lasso select the same factor (generates factor loading not equals to zero), and disregard the same factor (generate factor loadings equal to zero), we call Elastic net and Lasso made a exact agreement of factor selecting. We also extend our standard of the agreement to 90% interval. If the Elastic net and Lasso achieve agreement in 90% of factors' selection results, we would call they made a agreement on 90%. The proportion of agreement is presented in the table 6.3.

Table 6.3: Proportion of Lasso Regression and Elastic Net produces same or largely similar results for 145 companies

Factor Group	(0,0.5]	(0.5, 0.6]	(0.6, 0.7]	(0.7, 0.8]	(0.8,0.9]	(0.9,1]	Mix
Proportion of Agreement (Exact)	68.7%	55.9%	42.8%	20.9%	17.7%	13.9%	34.6%
Proportion of Agreement (90%)	86.8%	72.0%	74.5%	72.0%	79.8%	74.4%	76.1%

It is clear that the proportion of agreement, both exact agreed and 90% agreed, shows a decreasing trend with the factor strength increase. Nearly 70% of factors selection results are identical in the 0 to 0.5 strength group, but this number will decrease to 55% after we move to the 0.5-0.6 factor strength group. Only 14% of companies have identical factor selection results for group with strength between 0.9 and 1. For the mixed strength group, the exact agreed proportion is 34.6%, ranked between the 0.6 to 0.7 group and 0.7 to 0.8 group.

To provide an explanation of the disagreement increasing with factor strength increasing, we calculate the correlation among factors for each factor groups. The result is presented in table 6.4 and Figure B.3. We can clearly see an increasing trend of the correlation coefficient with

Table 6.4: Correlation Coefficient among different factor groups.

Factor Group	(0,0.5]	(0.5, 0.6]	(0.6, 0.7]	(0.7, 0.8]	(0.8,0.9]	(0.9,1]
Correlation Coefficient	0.0952	0.157	0.213	0.229	0.371	0.724
Factor Amount	12	10	17	37	35	34

the increase of factor's strength. This correlation pattern provides a possible explanation of the discord results of Elastic net and Lasso. When facing variables with high correlation, Lasso will randomly select several variables and discard others. While Elastic net address this problem

with the help of the extra L^2 norm in its loss function. Therefore, the Elastic net method will select all factors that can bring new information to explain the risk-return relationship even those factors are highly correlated. This also provides a potential explanation to what we observed in the table 5.3 that Lasso selects significantly fewer factors than the Elastic net method. Also, recall the fact that in the strong factor group, Elastic net will select two more factors than the Lasso on average. We believe Elastic net here will pick up factors that been abandoned by the Lasso due to the correlativity.

Chapter 7

Conclusion and Possible Extension

7.1 Conclusion

In this thesis, we propose the concept of factor strength and the corresponding estimation method. We applied the estimation on 145 different risk factors plus the market factor, estimating their strength and use the strength as reference to categorised each factors to reduce the dimension of potential factor group. On the basis of dimension-reduced factor group, we applied two feature selection methods namely Lasso and Elastic net, trying to eliminate the redundancy of factor groups.

From the factor strength estimation, we have a consistent estimation result of most of the strong factors. In general, factor strength will evolve in an upward manner with time increasing. And the strong factors will keep high strength through times. But we also notice that the overall factor strength is significantly lower when we use data set contains short time-spam, compare with results obtained from using larger data set with more observations. Also, it worth to notice that some conventionally strong factors, for instance, the size factor, value factor, and the momentum factor do not show particular high strength.

The empirical results of feature selection indicates that both Elastic net method and Lasso regression are both able to correctly and effectively eliminate factors without the ability to generating sufficient factor loadings. In another word, both Elastic net and Lasso can identify factor

with weak strength and disregard those redundant features. However, since the factors with strong strength are almost all highly correlated, the Elastic net and Lasso fail to make a mutual agreement when facing strong factors. On average, Elastic net will selects two more factors than Lasso when facing factors with strength above 0.9. We also noticed that when factors with different are mixed with each other, both lasso and elastic net can effectively pick up factor with strong strength and disregard weak factors.

7.2 Possible Extension

During the empirical exploration, we discovered some possible extension of this paper, and because of the time-limit as well as other restrictions, we could not propose those extensions. Here, we summarise those potential extensions to provide a starting point for future research.

Firstly, when discussing the factor strength, we ignore the heterogeneity of companies and factors. Ideally, companies and factors should be categorised based on their nature. For instance, companies from different industries may react differently to different factors. Therefore, one possible treatment is to group companies base on their industries types, or categorised factors base on their economical or financial meanings. Also, Harvey and Liu (2019) suggests that newly propsoed factors are more likely to encounter the multiple=testing problem, and hence they are more likely to be the factors that can not independently provides information to explain risk-return relationship. So, we could also categorised the factors base on their published time, and to investigates the relationship between factor strength and time.

Secondly, as discussed in the chapter 6.3,the principle of parameter tuning when applying elastic net is to select the parameter combination that minimise the MSE. However, the results present in the corresponding chapter indicates that MSE can not distinction different parameter combination significantly enough. Hence, considering use other criteria may lead to better performance of parameter tuning, and therefore improve the results of the Elastic net application.

Also, some other feature selection methods or dimension reduced techniques can be taken into consideration. The application of those methods can be used to cross-check with the re-

sults of Elastic and Lasso. Some potential methods including simple stepwise selection method, dantzig selector, and tree-based method like decision tree.

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

Dickey, D. A., & Fuller, W. A. (1979, 6). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, *74*, 427-431. Retrieved from https://www.tandfonline.com/doi/abs/10.1080/01621459.1979.10482531 doi: 10.1080/01621459.1979.10482531

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., & French, K. R. (2018, 5). Choosing factors. *Journal of Financial Economics*, *128*, 234-252. doi: 10.1016/j.jfineco.2018.02.012

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 non-

parametrically. *The Review of Financial Studies*, *33*, 2326-2377. Retrieved from https://doi.org/10.1093/rfs/hhz123 doi: 10.1093/rfs/hhz123

Friedman, J., Hastie, T., & Tibshirani, R. (2010, 2). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*, 1-22. Retrieved from https://www.jstatsoft.org/index.php/jss/article/view/v033i01/v33i01.pdf https://www.jstatsoft.org/index.php/jss/article/view/v033i01 doi: 10.18637/jss.v033.i01

Gospodinov, N., Kan, R., & Robotti, C. (2014, 7). Misspecification-robust inference in linear asset-pricing models with irrelevant risk factors. *Review of Financial Studies*, *27*, 2139-2170. Retrieved from https://academic.oup.com/rfs/article/27/7/2139/1578148 doi: 10.1093/rfs/hht135

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

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

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

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

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

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

Huang, J., Horowitz, J. L., & Wei, F. (2010, 8). Variable selection in nonparametric additive

models. *Annals of Statistics*, *38*, 2282-2313. doi: 10.1214/09-AOS781

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

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

Kleibergen, F., & 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

Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., & Shin, Y. (1992). *How sure are we that economic time series have a unit root?** (Vol. 54).

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*.

Phillips, P. C., & Perron, P. (1988, 6). Testing for a unit root in time series regression. *Biometrika*, *75*, 335-346. Retrieved from https://academic.oup.com/biomet/article/75/2/335/292919 doi: 10.1093/biomet/75.2.335

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

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2011, 3). Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, *39*, 1-13. Retrieved from https://www.jstatsoft.org/index.php/jss/article/view/v039i05/v39i05.pdf https://www.jstatsoft.org/index.php/jss/article/view/v039i05 doi: 10.18637/jss.v039.i05

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

Appendix A

Simulation Results

Table A.1: 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 A.2: Simulation result for double factors setting (no correlation)

	Double Factors with correlation $\rho_{12} = 0$								
	Bias × 100			RMSE × 100			Size × 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 A.3: Simulation result for double factors setting (weak correlation)

	Double Factors with correlation $\rho_{12} = 0.3$								
	Bias × 100			RMSE × 100			Size × 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 3. 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 $\alpha_1 = \alpha_2 = 1$

Table A.4: Simulation result for double factors setting (strong correlation)

	Double Factors with correlation $\rho_{12} = 0.7$								
	Bias $\times$ 100			RMSE $\times$ 100			Size $\times$ 100		
$\alpha_1 = 0.7$									
n\T	120	240	360	120	240	360	120	240	360
100	-0.659	0.069	0.059	1.193	0.406	0.372	46.95	2.65	1.25
300	-0.704	0.048	0.055	0.983	0.223	0.203	75.55	3.7	0.9
500	-0.841	-0.011	-0.004	1.056	0.161	0.144	86.4	5.1	0.3
$\alpha_1 = 0.75$									
100	-1.03	-0.315	-0.323	1.360	0.443	0.421	53.65	1.35	0.05
300	-0.724	0.049	0.065	0.947	0.166	0.159	84.75	4.85	0.6
500	-0.877	-0.055	-0.04	1.036	0.130	0.111	93.5	6.5	0.15
$\alpha_1 = 0.8$									
100	-1.099	-0.374	-0.362	1.36	0.436	0.421	94.6	86.05	85.75
300	-0.9	-0.114	-0.103	1.071	0.165	0.144	91.95	5.9	0.35
500	-0.822	0.005	0.015	0.968	0.086	0.072	97.3	9.2	0.8
$\alpha_1 = 0.85$									
100	-0.722	-0.001	0.009	1.006	0.175	0.157	71.75	3.00	0.65
300	-0.834	-0.05	-0.036	0.989	0.101	0.078	95.25	8.65	0.3
500	-0.883	-0.057	-0.039	1.013	0.084	0.062	98.95	13.3	0.15
$\alpha_1 = 0.9$									
100	-0.723	-0.004	0.001	0.972	0.125	0.107	77.35	2.8	0.5
300	-0.872	-0.055	-0.04	1.011	0.084	0.062	98.2	11.1	0.25
500	-0.851	-0.033	-0.018	0.967	0.06	0.037	99.75	17.7	0.75
$\alpha_1 = 0.95$									
100	-0.879	-0.116	-0.103	1.083	0.143	0.122	86.1	3.85	0.2
300	-0.853	-0.049	-0.035	0.977	0.066	0.044	99.55	14.65	1.2
500	-0.875	-0.029	-0.014	0.987	0.046	0.022	99.85	26.65	1.55
$\alpha_1 = 1$									
100	-0.76	-0.012	0	0.956	0054	0.009	-	-	-
300	-0.811	-0.015	0	0.945	0.037	0.004	-	-	-
500	-0.848	-0.017	0	0.96	0.033	0.003	-	-	-

Notes: This table shows the result of experiment 4. Factors and errors are generate from standard normal distribution. Between two factors, we assume they have correlation $\rho_{12} = 0.7$ 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 $\alpha_1 = \alpha_2 = 1$

Appendix B

Empirical Application Results

B.1 Empirical Factor Strength Estimates Tables

Table B.1: 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	0.995	0.937	salecash	0.998	0.948
2	baspread	0.980	0.730	quick	0.992	0.934	ndp	0.998	0.941
3	turn	0.983	0.728	salecash	0.993	0.933	quick	0.997	0.940
4	zerotrade	0.983	0.725	lev	0.994	0.931	age	0.997	0.940
5	idiovol	0.981	0.723	cash	0.993	0.931	roavol	0.997	0.938
6	retvol	0.978	0.721	dy	0.992	0.929	ep	0.998	0.937
7	std_turn	0.983	0.719	roavol	0.992	0.929	depr	0.998	0.935
8	HML_Devil	0.989	0.719	zs	0.994	0.927	cash	0.997	0.934
9	maxret	0.981	0.715	age	0.994	0.927	rds	0.998	0.931
10	roavol	0.986	0.713	cp	0.995	0.926	dy	0.997	0.927
11	age	0.989	0.703	ebp	0.994	0.926	currat	0.998	0.927

Table B.1: 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
12	sp	0.986	0.699	op	0.993	0.925	chcsho	0.997	0.927
13	ala	0.986	0.699	cfp	0.995	0.924	lev	0.996	0.926
14	ndp	0.988	0.686	nop	0.993	0.924	stdacc	0.997	0.926
15	orgcap	0.990	0.686	ep	0.994	0.923	cfp	0.998	0.925
16	tang	0.991	0.683	depr	0.993	0.922	nop	0.997	0.925
17	ebp	0.989	0.683	rds	0.993	0.922	zs	0.996	0.924
18	invest	0.986	0.683	kz	0.993	0.919	stdcf	0.997	0.924
19	dpia	0.987	0.681	sp	0.993	0.918	cp	0.998	0.919
20	UMD	0.990	0.678	currat	0.992	0.918	op	0.997	0.919
21	zs	0.987	0.675	ato	0.993	0.918	ato	0.996	0.919
22	grltnoa	0.989	0.675	chcsho	0.992	0.916	kz	0.996	0.918
23	dy	0.989	0.672	tang	0.995	0.915	ebp	0.997	0.918
24	HML	0.989	0.672	stdacc	0.992	0.913	tang	0.997	0.917
25	kz	0.987	0.669	adm	0.993	0.913	adm	0.997	0.913
26	ob_a	0.989	0.669	stdcf	0.991	0.910	ww	0.997	0.911
27	BAB	0.989	0.666	cashpr	0.994	0.909	maxret	0.995	0.911
28	op	0.991	0.663	nef	0.990	0.906	std_turn	0.995	0.908
29	realestate_hxz	0.988	0.663	HML	0.993	0.905	idiovol	0.994	0.908
30	ol	0.988	0.663	std_turn	0.991	0.901	nef	0.995	0.908
31	adm	0.989	0.660	idiovol	0.990	0.901	baspread	0.995	0.906
32	lev	0.987	0.657	zerotrade	0.988	0.897	IPO	0.997	0.902
33	nxf	0.990	0.651	ww	0.994	0.896	retvol	0.995	0.902
34	nop	0.990	0.651	turn	0.989	0.895	sp	0.995	0.901
35	pm	0.986	0.648	maxret	0.990	0.893	turn	0.994	0.900
36	pchcapx3	0.989	0.644	absacc	0.995	0.889	absacc	0.998	0.898
37	nef	0.989	0.644	baspread	0.990	0.885	lgr	0.997	0.897
38	cash	0.990	0.637	hire	0.994	0.883	zerotrade	0.992	0.896
39	QMJ	0.978	0.637	lgr	0.994	0.882	HML	0.995	0.894
40	rds	0.989	0.634	IPO	0.994	0.881	cashpr	0.995	0.892
41	LIQ_PS	0.989	0.634	retvol	0.990	0.879	salerec	0.997	0.890
42	ato	0.989	0.634	nxf	0.990	0.879	dcol	0.997	0.890

Table B.1: 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
43	salerec	0.992	0.630	RMW	0.992	0.879	hire	0.997	0.890
44	currat	0.989	0.626	beta	0.988	0.878	RMW	0.997	0.890
45	acc	0.990	0.619	salerec	0.992	0.877	beta	0.993	0.889
46	stdcf	0.990	0.619	acc	0.995	0.875	nxf	0.996	0.886
47	HXZ_ROE	0.989	0.619	bm_ia	0.995	0.875	acc	0.997	0.882
48	depr	0.989	0.615	sin	0.994	0.874	dfin	0.996	0.875
49	noa	0.989	0.615	dcol	0.994	0.872	nincr	0.996	0.872
50	cashpr	0.988	0.615	dfin	0.993	0.870	noa	0.995	0.868
51	absacc	0.990	0.615	HML_Devil	0.988	0.870	HXZ_IA	0.997	0.868
52	gma	0.988	0.615	HXZ_IA	0.994	0.869	rdm	0.996	0.867
53	dncl	0.987	0.611	nincr	0.994	0.864	HML_Devil	0.995	0.867
54	ms	0.981	0.611	rdm	0.993	0.855	rna	0.997	0.865
55	rna	0.990	0.611	noa	0.992	0.855	ps	0.996	0.857
56	STR	0.988	0.607	rna	0.994	0.855	bm_ia	0.997	0.856
57	rdm	0.989	0.607	herf	0.991	0.854	sgr	0.997	0.854
58	chcsho	0.987	0.607	sgr	0.993	0.849	rd	0.996	0.852
59	sin	0.988	0.607	dnco	0.993	0.845	sin	0.997	0.852
60	salecash	0.989	0.602	ps	0.992	0.836	realestate_hxz	0.997	0.851
61	dnco	0.989	0.598	CMA	0.994	0.836	herf	0.995	0.846
62	quick	0.990	0.593	egr_hxz	0.992	0.832	dnco	0.996	0.844
63	stdacc	0.990	0.593	realestate_hxz	0.991	0.830	CMA	0.997	0.838
64	poa	0.989	0.593	rd	0.993	0.820	egr_hxz	0.996	0.831
65	cp	0.989	0.589	cinvest_a	0.994	0.817	ol	0.995	0.829
66	tb	0.989	0.589	ol	0.989	0.816	ob_a	0.996	0.823
67	HXZ_IA	0.988	0.584	gad	0.992	0.816	cinvest_a	0.996	0.823
68	saleinv	0.987	0.579	dolvol	0.995	0.804	SMB	0.995	0.804
69	cfp	0.989	0.579	ob_a	0.990	0.797	gad	0.995	0.804
70	egr	0.988	0.579	pchdepr	0.994	0.791	dolvol	0.997	0.798
71	dnca	0.987	0.579	ala	0.993	0.791	gma	0.995	0.798
72	egr_hxz	0.989	0.579	BAB	0.995	0.785	ala	0.996	0.798
73	os	0.985	0.569	pchcapx3	0.991	0.782	QMJ	0.996	0.795

Table B.1: 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
74	pps	0.983	0.563	gma	0.990	0.782	convind	0.997	0.793
75	cto	0.988	0.563	dnca	0.992	0.780	tb	0.995	0.791
76	grltnoa_hxz	0.987	0.563	SMB	0.992	0.771	aeavol	0.998	0.791
77	cei	0.989	0.563	poa	0.991	0.769	BAB	0.998	0.788
78	CMA	0.989	0.563	aeavol	0.996	0.767	dcoa	0.995	0.784
79	em	0.989	0.552	tb	0.988	0.763	cto	0.995	0.781
80	ww	0.991	0.546	grltnoa_hxz	0.992	0.761	egr	0.996	0.781
81	std_dolvol	0.988	0.539	cei	0.988	0.761	roic	0.995	0.781
82	grcapx	0.987	0.539	dsti	0.991	0.757	indmom	0.995	0.779
83	pctacc	0.989	0.539	indmom	0.990	0.755	pm	0.995	0.779
84	ep	0.989	0.533	egr	0.992	0.753	pchdepr	0.996	0.773
85	pricedelay	0.989	0.533	moms12m	0.992	0.753	cei	0.995	0.773
86	hire	0.989	0.519	orgcap	0.990	0.744	orgcap	0.995	0.771
87	SMB	0.988	0.512	dcoa	0.993	0.737	pricedelay	0.996	0.768
88	pchcapx_ia	0.989	0.512	pchcurrat	0.993	0.735	dnca	0.995	0.768
89	aeavol	0.989	0.512	UMD	0.986	0.733	moms12m	0.995	0.768
90	moms12m	0.988	0.512	cinvest	0.993	0.733	pchcapx3	0.995	0.765
91	cashdebt	0.985	0.504	HXZ_ROE	0.992	0.733	saleinv	0.995	0.763
92	lgr	0.988	0.504	roic	0.986	0.730	pctacc	0.995	0.763
93	cinvest	0.989	0.496	QMJ	0.985	0.730	grltnoa_hxz	0.996	0.760
94	herf	0.988	0.496	pctacc	0.989	0.723	poa	0.995	0.757
95	bm_ia	0.989	0.487	cto	0.990	0.718	HXZ_ROE	0.997	0.757
96	cfp_ia	0.988	0.479	pricedelay	0.993	0.715	UMD	0.995	0.745
97	cinvest_a	0.989	0.479	pchcapx_ia	0.992	0.707	dsti	0.995	0.742
98	chmom	0.990	0.469	convind	0.990	0.698	cinvest	0.995	0.733
99	RMW	0.988	0.469	cdi	0.993	0.677	em	0.995	0.733
100	sue	0.988	0.459	invest	0.991	0.677	pchcurrat	0.995	0.730
101	mom36m	0.987	0.459	chtx	0.992	0.674	ms	0.995	0.716
102	indmom	0.988	0.459	rsup	0.991	0.670	pchcapx_ia	0.995	0.708
103	dcoa	0.989	0.459	em	0.987	0.670	invest	0.995	0.708
104	etr	0.987	0.448	pm	0.992	0.667	dpia	0.995	0.705

Table B.1: 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
105	chinv	0.989	0.448	ta	0.992	0.663	os	0.992	0.701
106	ill	0.989	0.448	dpia	0.991	0.663	cdi	0.995	0.701
107	roic	0.987	0.448	saleinv	0.991	0.660	chtx	0.995	0.689
108	convind	0.989	0.448	pchquick	0.991	0.652	pps	0.995	0.680
109	sgr	0.989	0.437	os	0.983	0.652	rs	0.995	0.680
110	IPO	0.990	0.437	ms	0.985	0.640	roaq	0.995	0.680
111	dolvol	0.990	0.437	roaq	0.988	0.628	rsup	0.995	0.642
112	dcol	0.988	0.425	pps	0.986	0.623	cfp_ia	0.995	0.637
113	nincr	0.989	0.411	cfp_ia	0.991	0.623	chinv	0.995	0.637
114	chempia	0.988	0.411	grcapx	0.990	0.619	ta	0.995	0.631
115	rs	0.989	0.411	ndf	0.991	0.614	cashdebt	0.993	0.625
116	pchcapx	0.989	0.411	mom6m	0.992	0.604	STR	0.995	0.619
117	chtx	0.989	0.397	dncl	0.991	0.604	ndf	0.995	0.619
118	ivg	0.989	0.381	pchsale_pchrect	0.990	0.599	grltnoa	0.995	0.613
119	LTR	0.986	0.364	pchcapx	0.992	0.599	pchcapx	0.995	0.613
120	mom6m	0.988	0.364	pchsaleinv	0.990	0.594	pchquick	0.995	0.607
121	cdi	0.988	0.364	LIQ_PS	0.990	0.588	mom6m	0.995	0.607
122	chatoia	0.988	0.364	rs	0.990	0.588	grcapx	0.995	0.607
123	gad	0.986	0.364	cashdebt	0.986	0.583	dncl	0.995	0.600
124	pchcurrat	0.989	0.297	chempia	0.992	0.583	ivg	0.995	0.586
125	pchgm_pchsale	0.989	0.297	dwc	0.990	0.565	pchsaleinv	0.995	0.579
126	rd	0.987	0.297	grltnoa	0.990	0.551	LIQ_PS	0.995	0.579
127	dsti	0.990	0.297	dfnl	0.990	0.544	chempia	0.995	0.554
128	dfnl	0.988	0.297	STR	0.990	0.537	mom36m	0.995	0.545
129	roaq	0.986	0.297	std_dolvol	0.990	0.537	pchsale_pchxsga	0.995	0.526
130	pchdepr	0.989	0.266	mom36m	0.991	0.513	std_dolvol	0.995	0.526
131	dnoa	0.989	0.230	sue	0.990	0.504	pchsale_pchinvt	0.995	0.516
132	ta	0.989	0.230	LTR	0.989	0.504	dwc	0.995	0.516
133	chpmia	0.988	0.230	chmom	0.988	0.504	chmom	0.994	0.505
134	pchquick	0.988	0.182	pchsale_pchxsga	0.991	0.475	sue	0.995	0.493
135	dwc	0.990	0.182	pchsale_pchinvt	0.990	0.464	dfnl	0.995	0.480

Table B.1: 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
136	dfin	0.989	0.182	lfe	0.990	0.452	LTR	0.995	0.467
137	rsup	0.989	0.182	chinv	0.991	0.439	pchsale_pchrect	0.995	0.467
138	pchsaleinv	0.989	0.115	chatoia	0.991	0.439	pchgm_pchsale	0.995	0.379
139	pchsale_pchinvt	0.989	0.115	ivg	0.991	0.426	lfe	0.995	0.379
140	pchsale_pchrect	0.989	0.115	pchgm_pchsale	0.991	0.394	ill	0.995	0.326
141	pchsale_pchxsga	0.990	0.115	etr	0.990	0.356	dnoa	0.995	0.293
142	ps	0.991	0.115	chpmia	0.991	0.356	ear	0.995	0.252
143	lfe	0.989	0.000	ill	0.990	0.276	etr	0.995	0.200
144	ndf	0.987	0.000	dnoa	0.990	0.276	chatoia	0.995	0.200
145	ear	0.989	0.000	ear	0.992	0.276	chpmia	0.995	0.200

Notes: This table presents the estimation results of factors' strength, ordered decreasingly by risk factor strength. For the estimation, we use the method from chapter 3.3, with one market factor and one risk factor. The three data set is describe in the chapter 5.1

Table B.2: 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	0.763	ndp	0.937	salecash	0.948
2	saleinv	0.736	quick	0.934	ndp	0.941
3	cto	0.733	salecash	0.933	quick	0.940
4	turn	0.733	lev	0.931	age	0.940
5	beta	0.730	cash	0.931	roavol	0.938
6	ala	0.730	dy	0.929	ep	0.937
7	nef	0.726	roavol	0.929	depr	0.935
8	zerotrade	0.719	zs	0.927	cash	0.934
9	dy	0.716	age	0.927	rds	0.931
10	idiovol	0.716	cp	0.926	dy	0.927
11	ol	0.712	ebp	0.926	currat	0.927
12	gma	0.712	op	0.925	chcsho	0.927
13	std_turn	0.708	cfp	0.924	lev	0.926
14	depr	0.705	nop	0.924	stdacc	0.926
15	baspread	0.701	ep	0.923	cfp	0.925
16	retvol	0.701	depr	0.922	nop	0.925
17	nxf	0.701	rds	0.922	zs	0.924
18	currat	0.697	kz	0.919	stdcf	0.924
19	maxret	0.697	sp	0.918	cp	0.919
20	op	0.693	currat	0.918	op	0.919
21	SMB	0.689	ato	0.918	ato	0.919
22	orgcap	0.685	chcsho	0.916	kz	0.918
23	nop	0.680	tang	0.915	ebp	0.918
24	pm	0.680	stdacc	0.913	tang	0.917
25	tang	0.676	adm	0.913	adm	0.913
26	quick	0.672	stdcf	0.910	ww	0.911
27	roavol	0.667	cashpr	0.909	maxret	0.911
28	pricedelay	0.667	nef	0.906	std_turn	0.908
29	sp	0.657	HML	0.905	idiovol	0.908
30	aeavol	0.657	std_turn	0.901	nef	0.908
31	bm_ia	0.652	idiovol	0.901	baspread	0.906
32	cash	0.652	zerotrade	0.897	IPO	0.902
33	convind	0.652	ww	0.896	retvol	0.902
34	dcoa	0.642	turn	0.895	sp	0.901
35	ebp	0.642	maxret	0.893	turn	0.900
36	ivg	0.642	absacc	0.889	absacc	0.898
37	cp	0.637	baspread	0.885	lgr	0.897
38	chinv	0.637	hire	0.883	zerotrade	0.896
39	ndp	0.637	lgr	0.882	HML	0.894
40	hire	0.637	IPO	0.881	cashpr	0.892
41	roic	0.631	retvol	0.879	salerec	0.890
42	cashpr	0.625	nxf	0.879	dcol	0.890
43	HML_Devil	0.625	RMW	0.879	hire	0.890
44	HXZ_IA	0.625	beta	0.878	RMW	0.890
45	HML	0.619	salerec	0.877	beta	0.889
46	age	0.619	acc	0.875	nxf	0.886
47	egr_hxz	0.619	bm_ia	0.875	acc	0.882
48	dpia	0.619	sin	0.874	dfin	0.875
49	invest	0.619	dcol	0.872	nincr	0.872
50	poa	0.619	dfin	0.870	noa	0.868
51	QMJ	0.619	HML_Devil	0.870	HXZ_IA	0.868
52	salerec	0.613	HXZ_IA	0.869	rdm	0.867

Table B.2: 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
53	dnco	0.613	nincr	0.864	HML_Devil	0.867
54	cdi	0.613	rdm	0.855	rna	0.865
55	em	0.613	noa	0.855	ps	0.857
56	salecash	0.607	rna	0.855	bm_ia	0.856
57	sgr	0.607	herf	0.854	sgr	0.854
58	egr	0.607	sgr	0.849	rd	0.852
59	dcol	0.607	dnco	0.845	sin	0.852
60	pchcapx3	0.600	ps	0.836	realestate_hxz	0.851
61	kz	0.593	CMA	0.836	herf	0.846
62	lev	0.586	egr_hxz	0.832	dnco	0.844
63	acc	0.586	realestate_hxz	0.830	CMA	0.838
64	zs	0.586	rd	0.820	egr_hxz	0.831
65	rsup	0.586	cinvest_a	0.817	ol	0.829
66	pps	0.579	ol	0.816	ob_a	0.823
67	nincr	0.579	gad	0.816	cinvest_a	0.823
68	rdm	0.579	dolvol	0.804	SMB	0.804
69	grltnoa_hxz	0.579	ob_a	0.797	gad	0.804
70	cfp	0.579	pchdepr	0.791	dolvol	0.798
71	chcsho	0.579	ala	0.791	gma	0.798
72	pctacc	0.571	BAB	0.785	ala	0.798
73	CMA	0.571	pchcapx3	0.782	QMJ	0.795
74	ep	0.563	gma	0.782	convind	0.793
75	UMD	0.563	dnca	0.780	tb	0.791
76	indmom	0.563	SMB	0.771	aeavol	0.791
77	dwc	0.563	poa	0.769	BAB	0.788
78	dnca	0.563	aeavol	0.767	dcoa	0.784
79	chmom	0.563	tb	0.763	cto	0.781
80	cei	0.563	grltnoa_hxz	0.761	egr	0.781
81	lgr	0.545	cei	0.761	roic	0.781
82	STR	0.536	dsti	0.757	indmom	0.779
83	absacc	0.536	indmom	0.755	pm	0.779
84	realestate_hxz	0.536	egr	0.753	pchdepr	0.773
85	chempia	0.505	moms12m	0.753	cei	0.773
86	rds	0.505	orgcap	0.744	orgcap	0.771
87	ps	0.493	dcoa	0.737	pricedelay	0.768
88	grltnoa	0.493	pchcurrat	0.735	dnca	0.768
89	ms	0.493	UMD	0.733	moms12m	0.768
90	rs	0.493	cinvest	0.733	pchcapx3	0.765
91	RMW	0.493	HXZ_ROE	0.733	saleinv	0.763
92	os	0.480	roic	0.730	pctacc	0.763
93	ob_a	0.480	QMJ	0.730	grltnoa_hxz	0.760
94	roaq	0.480	pctacc	0.723	poa	0.757
95	cinvest	0.467	cto	0.718	HXZ_ROE	0.757
96	ww	0.467	pricedelay	0.715	UMD	0.745
97	BAB	0.467	pchcapx_ia	0.707	dsti	0.742
98	dolvol	0.452	convind	0.698	cinvest	0.733
99	std_dolvol	0.452	cdi	0.677	em	0.733
100	rd	0.452	invest	0.677	pchcurrat	0.730
101	grcapx	0.452	chtx	0.674	ms	0.716
102	moms12m	0.452	rsup	0.670	pchcapx_ia	0.708
103	chatoia	0.452	em	0.670	invest	0.708
104	mom36m	0.437	pm	0.667	dpia	0.705
105	chtx	0.419	ta	0.663	os	0.701

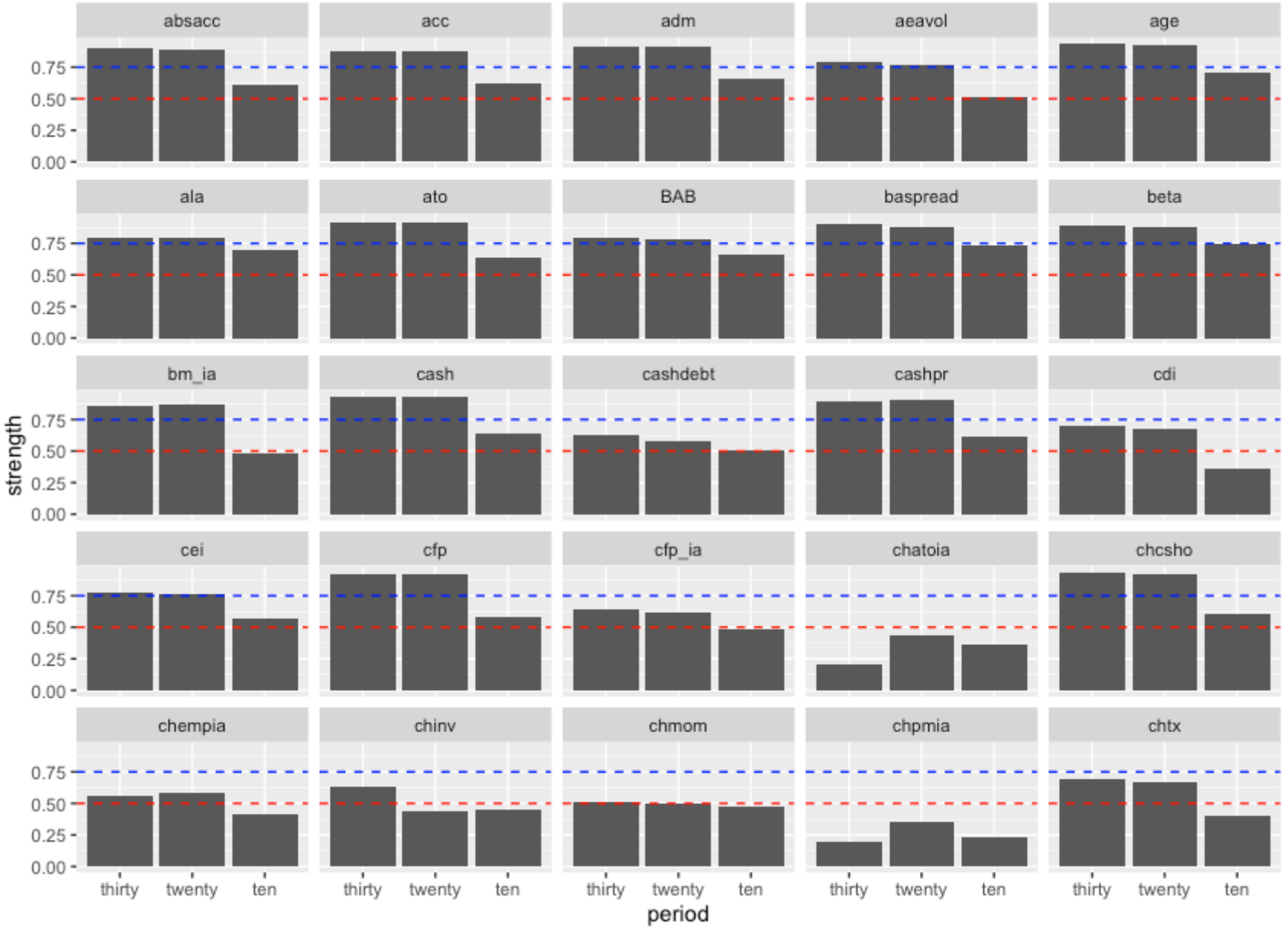
Table B.2: Decompose the thirty year data into three ten year subset, estimated the factor strength base on those three data set separately(cont.)

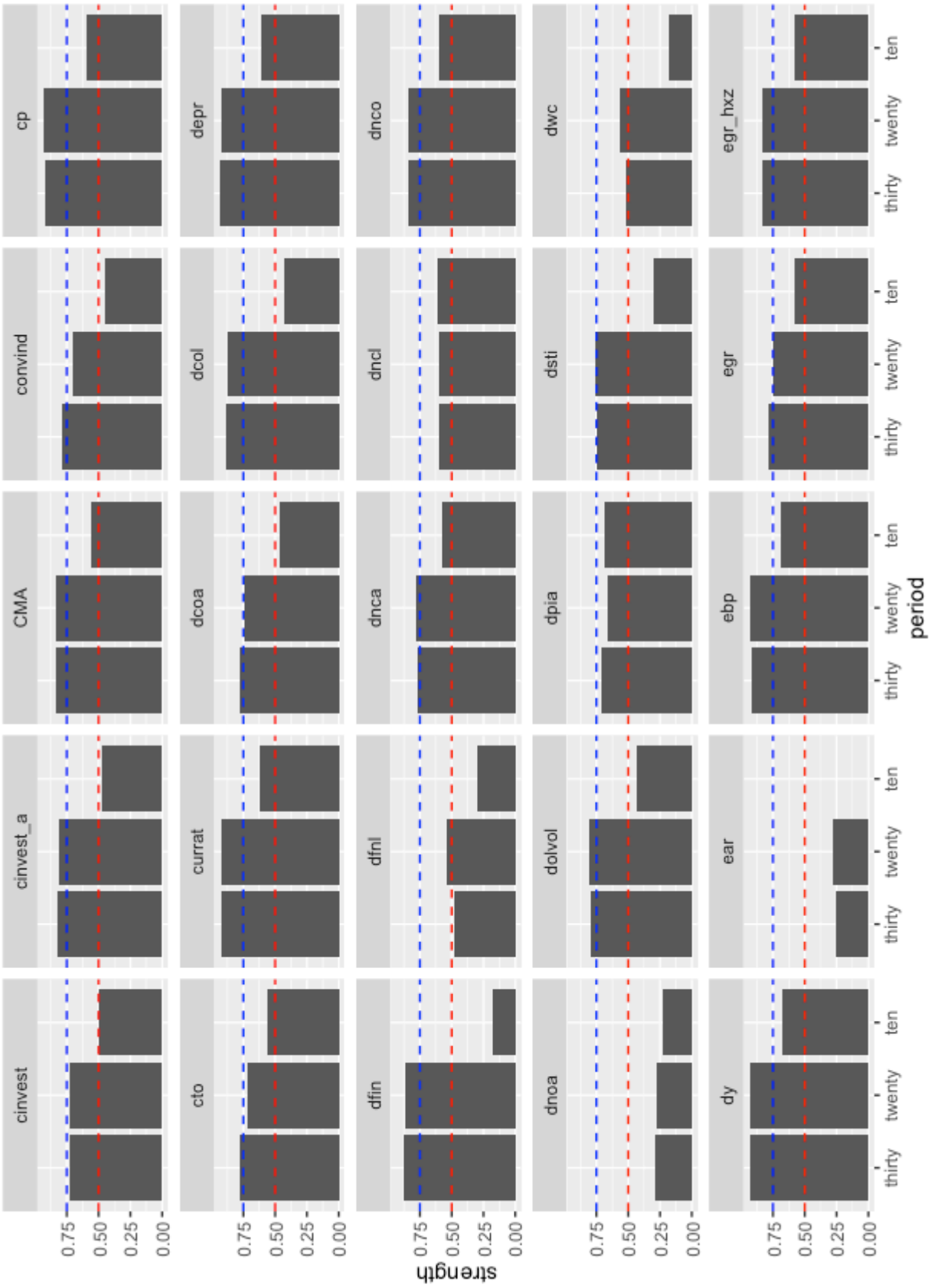
	January 1988 to December 1997		January 1998 to December 2007		January 2008 to December 2017	
Rank	Factor	Strength	Factor	Strength	Factor	Strength
106	ato	0.419	dpia	0.663	cdi	0.701
107	stdcf	0.419	saleinv	0.660	chtx	0.689
108	cashdebt	0.400	pchquick	0.652	pps	0.680
109	ta	0.400	os	0.652	rs	0.680
110	stdacc	0.400	ms	0.640	roaq	0.680
111	HXZ_ROE	0.400	roaq	0.628	rsup	0.642
112	IPO	0.379	pps	0.623	cfp_ia	0.637
113	cfp_ia	0.379	cfp_ia	0.623	chinv	0.637
114	dfin	0.379	grcapx	0.619	ta	0.631
115	dfnl	0.379	ndf	0.614	cashdebt	0.625
116	pchcapx	0.379	mom6m	0.604	STR	0.619
117	adm	0.354	dncl	0.604	ndf	0.619
118	noa	0.354	pchsale_pchrect	0.599	grltnoa	0.613
119	rna	0.326	pchcapx	0.599	pchcapx	0.613
120	pchsale_pchinvt	0.293	pchsaleinv	0.594	pchquick	0.607
121	pchsale_pchxsga	0.293	LIQ_PS	0.588	mom6m	0.607
122	etr	0.293	rs	0.588	grcapx	0.607
123	lfe	0.293	cashdebt	0.583	dncl	0.600
124	cinvest_a	0.293	chempia	0.583	ivg	0.586
125	ndf	0.293	dwc	0.565	pchsaleinv	0.579
126	sue	0.252	grltnoa	0.551	LIQ_PS	0.579
127	gad	0.252	dfnl	0.544	chempia	0.554
128	LTR	0.200	STR	0.537	mom36m	0.545
129	pchsaleinv	0.200	std_dolvol	0.537	pchsale_pchxsga	0.526
130	mom6m	0.200	mom36m	0.513	std_dolvol	0.526
131	ill	0.200	sue	0.504	pchsale_pchinvt	0.516
132	LIQ_PS	0.200	LTR	0.504	dwc	0.516
133	tb	0.200	chmom	0.504	chmom	0.505
134	sin	0.200	pchsale_pchxsga	0.475	sue	0.493
135	pchcurrat	0.126	pchsale_pchinvt	0.464	dfnl	0.480
136	pchsale_pchrect	0.126	lfe	0.452	LTR	0.467
137	pchcapx_ia	0.126	chinv	0.439	pchsale_pchrect	0.467
138	dncl	0.126	chatoia	0.439	pchgm_pchsale	0.379
139	dsti	0.126	ivg	0.426	lfe	0.379
140	ear	0.126	pchgm_pchsale	0.394	ill	0.326
141	pchquick	0.000	etr	0.356	dnoa	0.293
142	pchdepr	0.000	chpmia	0.356	ear	0.252
143	pchgm_pchsale	0.000	ill	0.276	etr	0.200
144	dnoa	0.000	dnoa	0.276	chatoia	0.200
145	chpmia	0.000	ear	0.276	chpmia	0.200

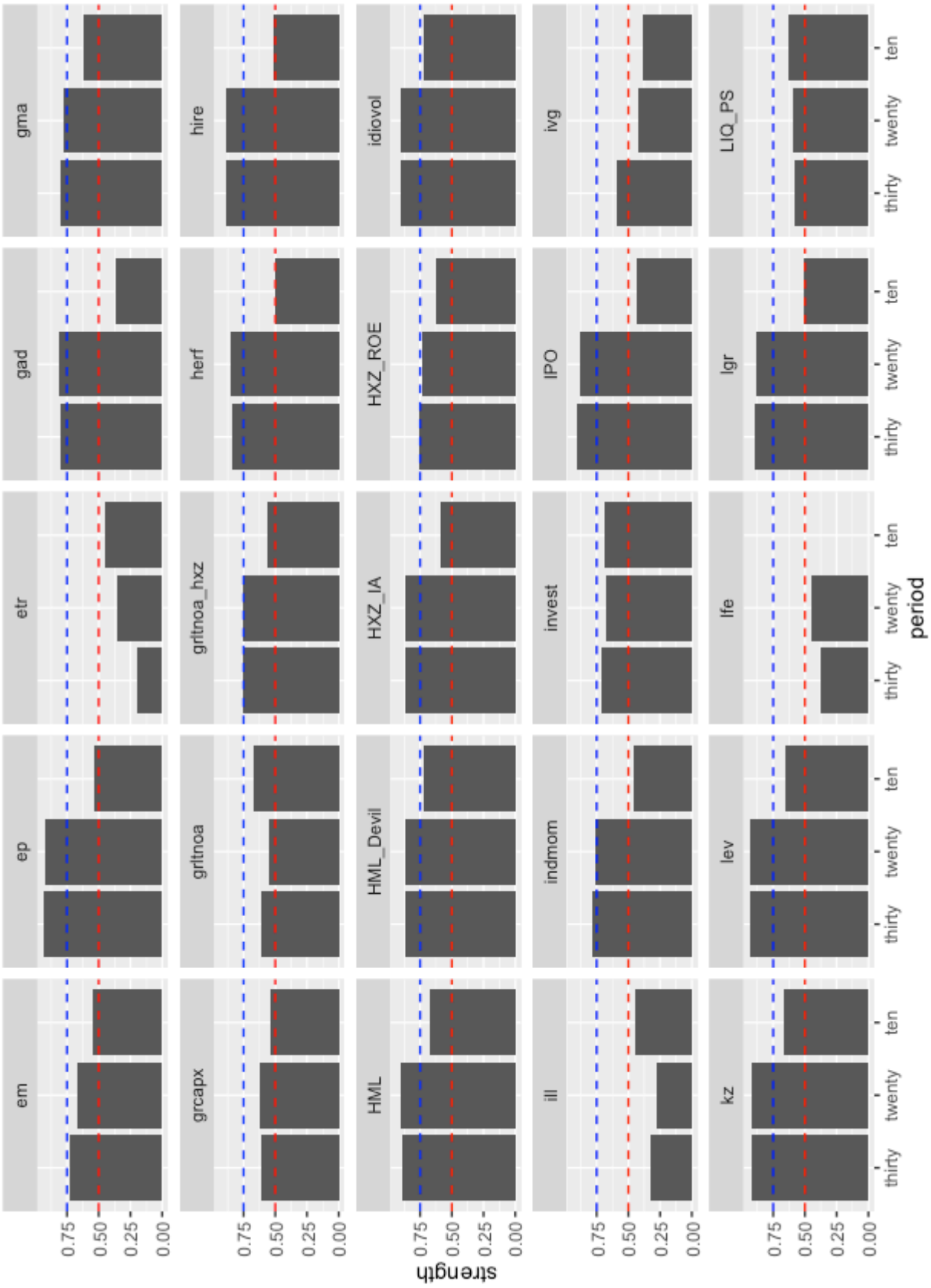
Notes: This table presents the estimated factor strength, using the decomposed thirty years data. The thirty year data set is decomposed into three subsets: January 1988 to December 1997, January 1998 to December 2007, and January 2008 to December 2017. For each data set, it contains 120 observations (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.

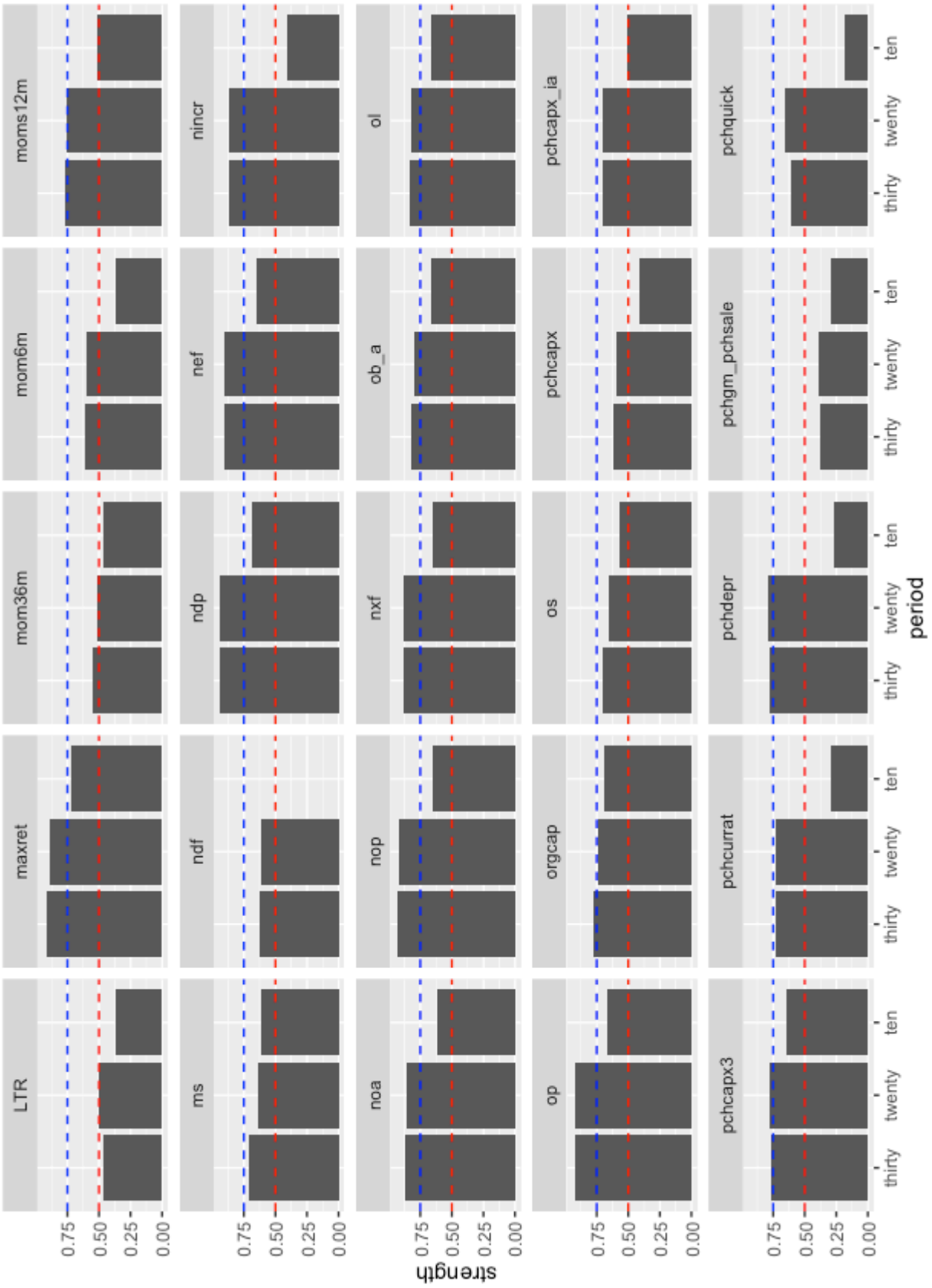
B.2 Strength Comparisons Figures

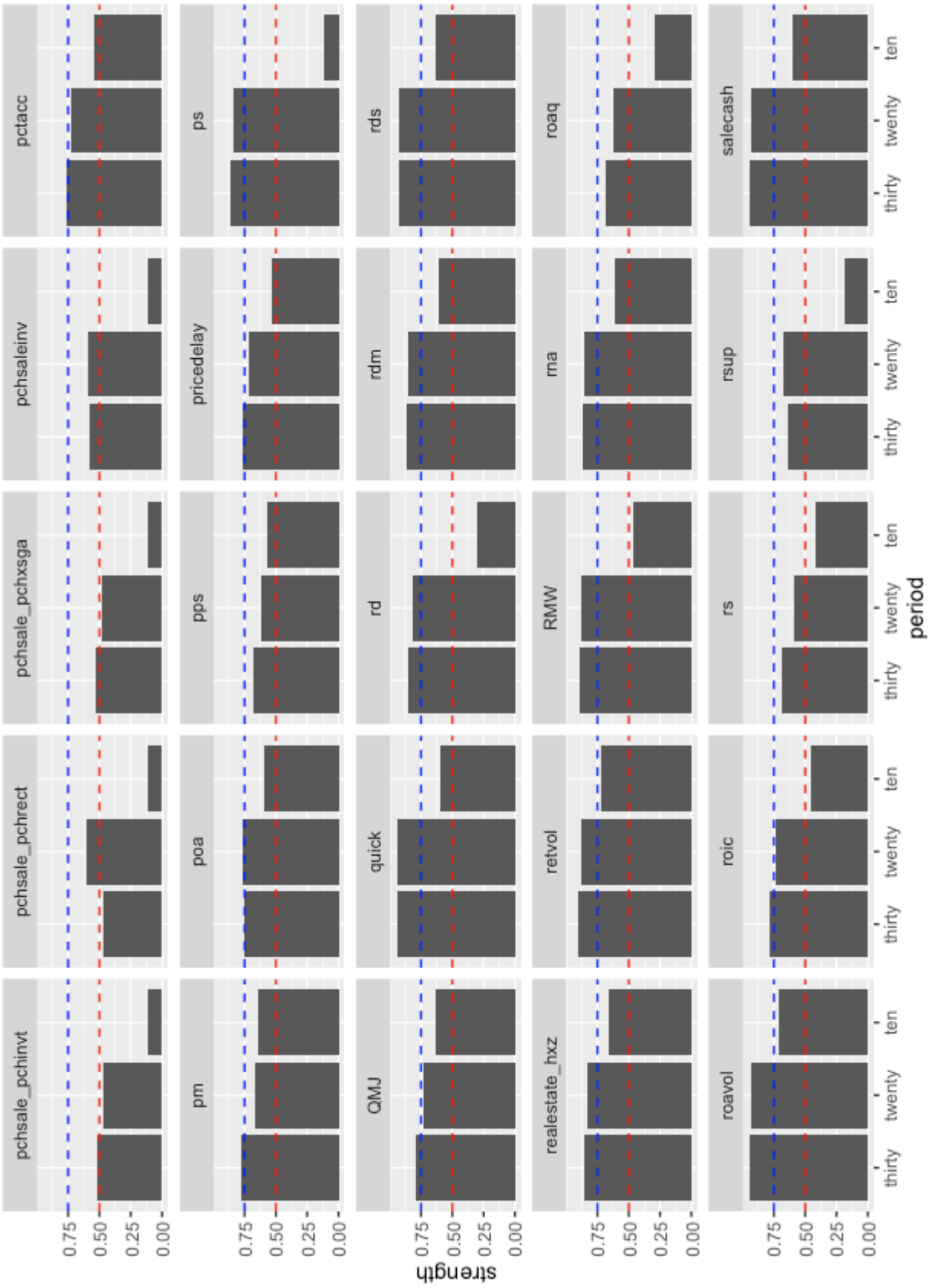
Figure B.1: Strength Comparison

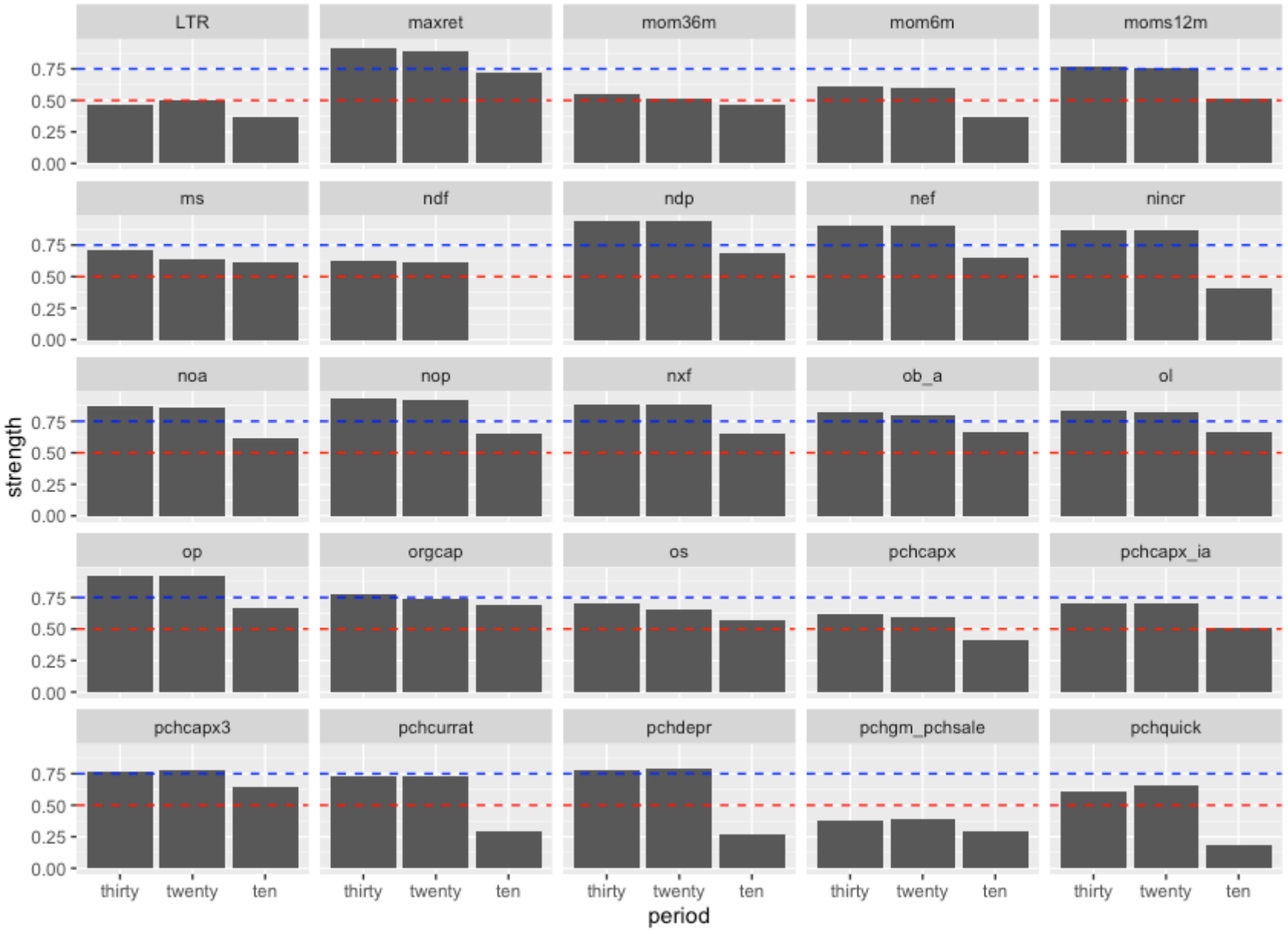



cp
convind
CMA
cinvest_a
cinvest
depr
dcol
dcoa
currat
cto
dnco
dncl
dnca
dfnl
dfin
dwc
dsti
dpia
dolvol
dnoa
egr_hxz
egr
ebp
ear
dy
0.75
0.50
0.25
0.00
strength
thirty
twenty
ten
period


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

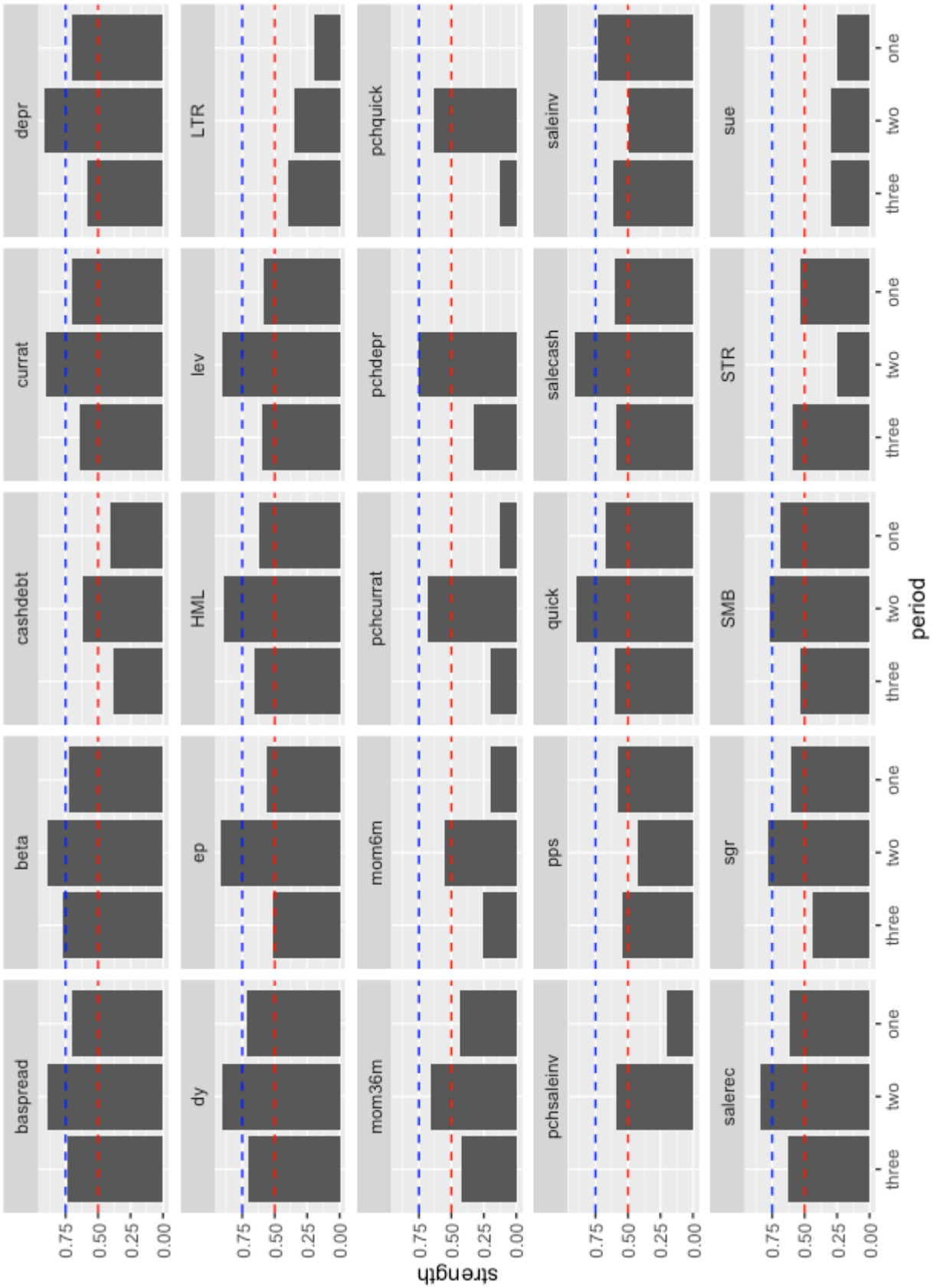

moms12m
nincr
ol
pchcapx_ia
pchquick
mom6m
nef
ob_a
pchcapx
pchgm_pchsale
mom36m
ndp
nxf
os
pchdepr
maxret
ndf
nop
orgcap
pchcurrat
LTR
ms
noa
op
pchcapx3
0.75
0.50
0.25
0.00
strength
thirty
twenty
ten
period

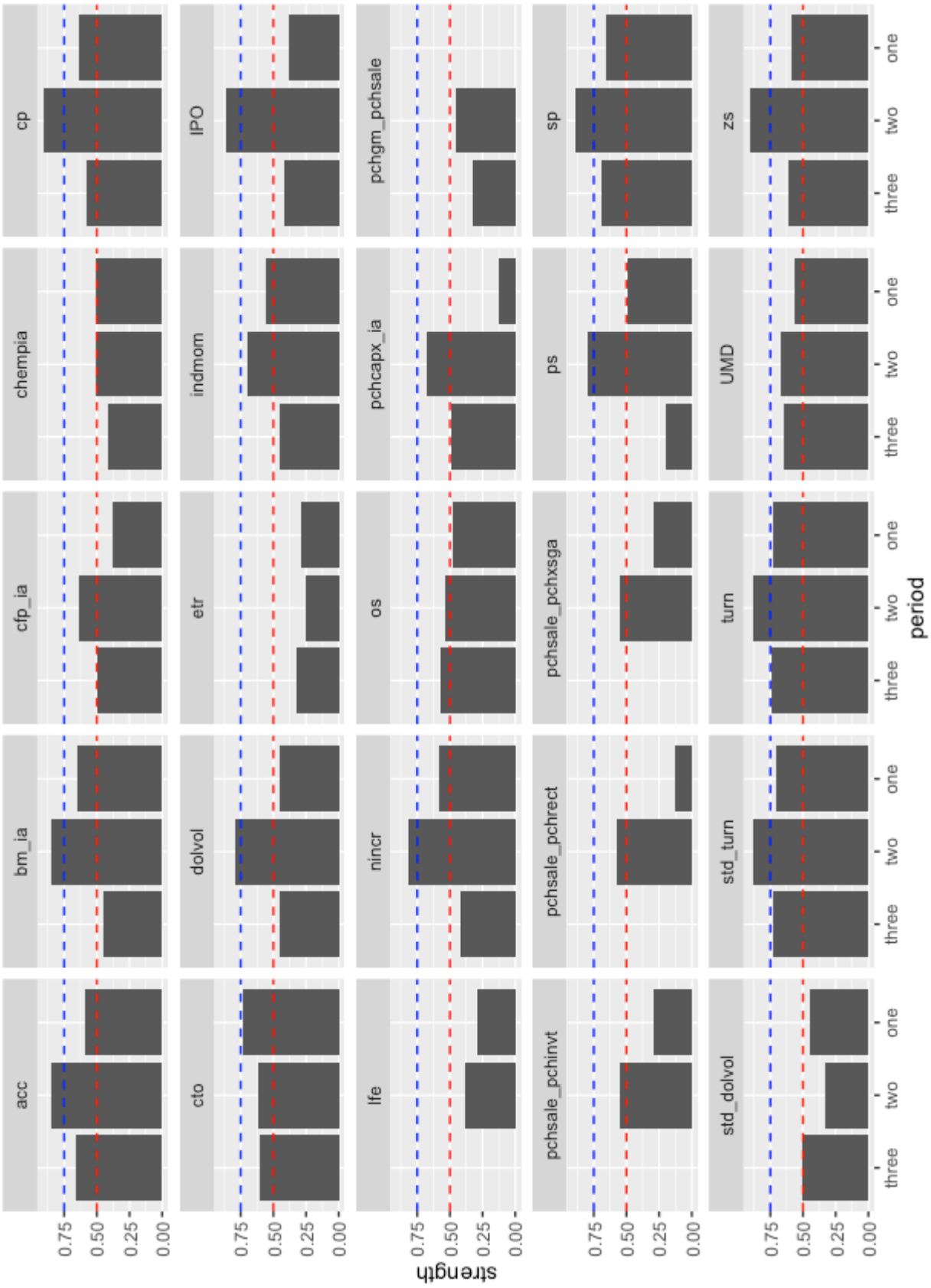

pctacc
pchsaleinv
pchsale_pchxsga
pchsale_pchrect
pchsale_pchinvt
ps
pricedelay
pps
poa
pm
rds
rdm
rd
quick
QMJ
roaq
rna
RMW
retvol
realestate_hxz
salecash
rsup
rs
roic
roavol
0.75
0.50
0.25
0.00
strength
thirty
twenty
ten
period

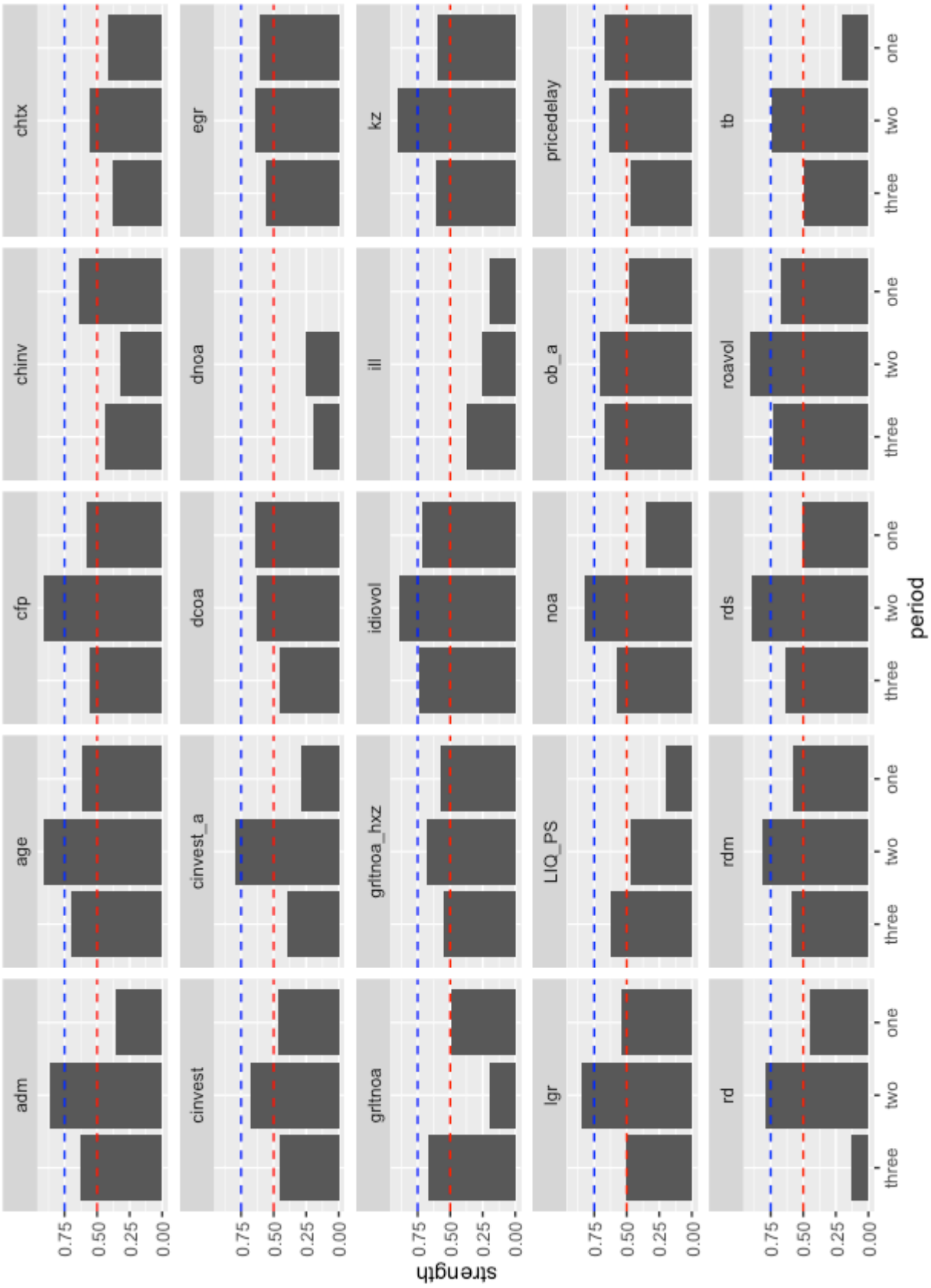


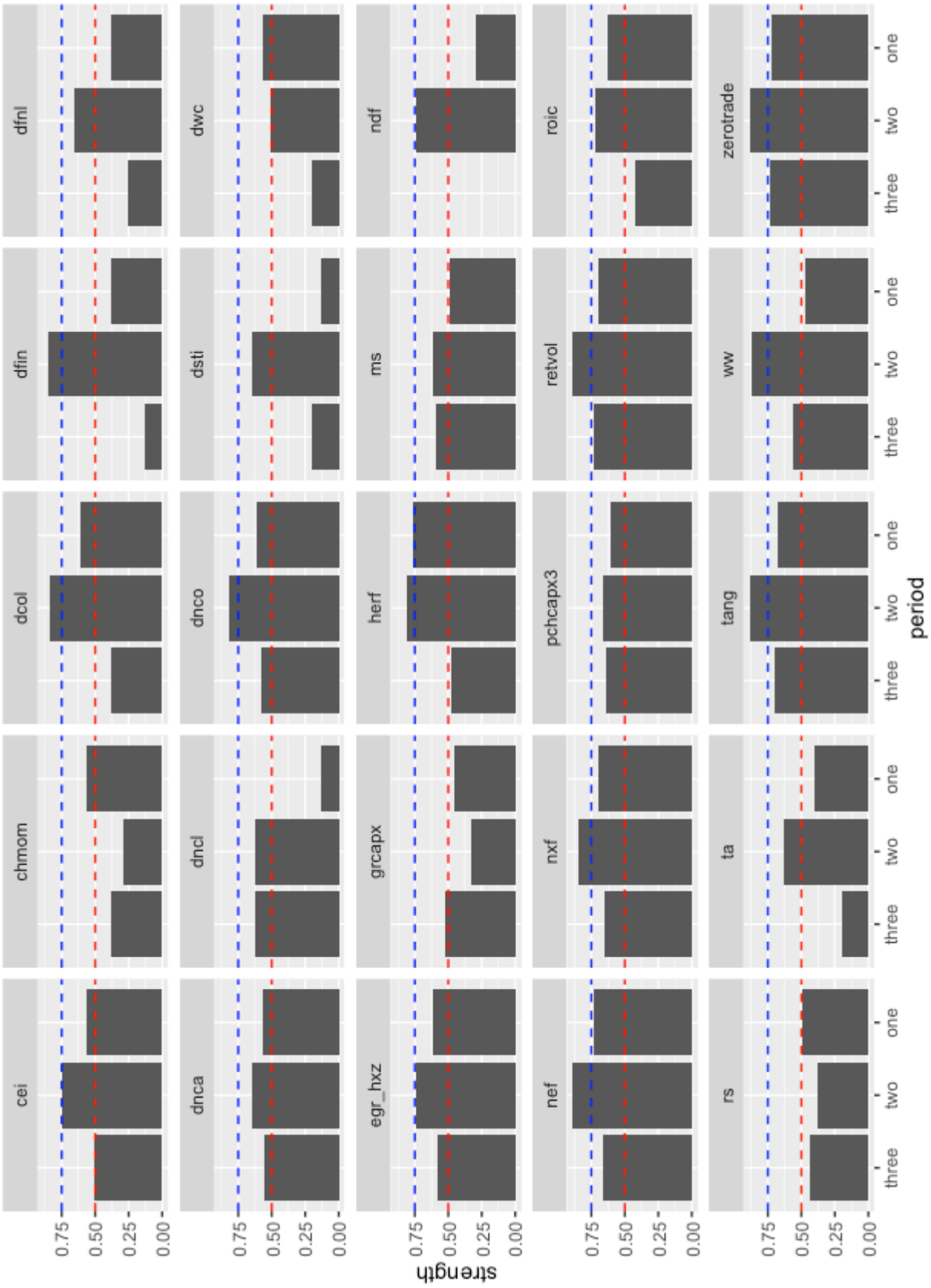
Notes: The figure compare the strength of every factor's strength in different data set. The x-axis indicates the data set: thirty is thirty years data set (January 1987 to December 2017), twenty is twenty year data set (January 1997 to December 2017), and ten is ten year data set (January 2007 to December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

Figure B.2: Thirty Year Decomposition Comparison

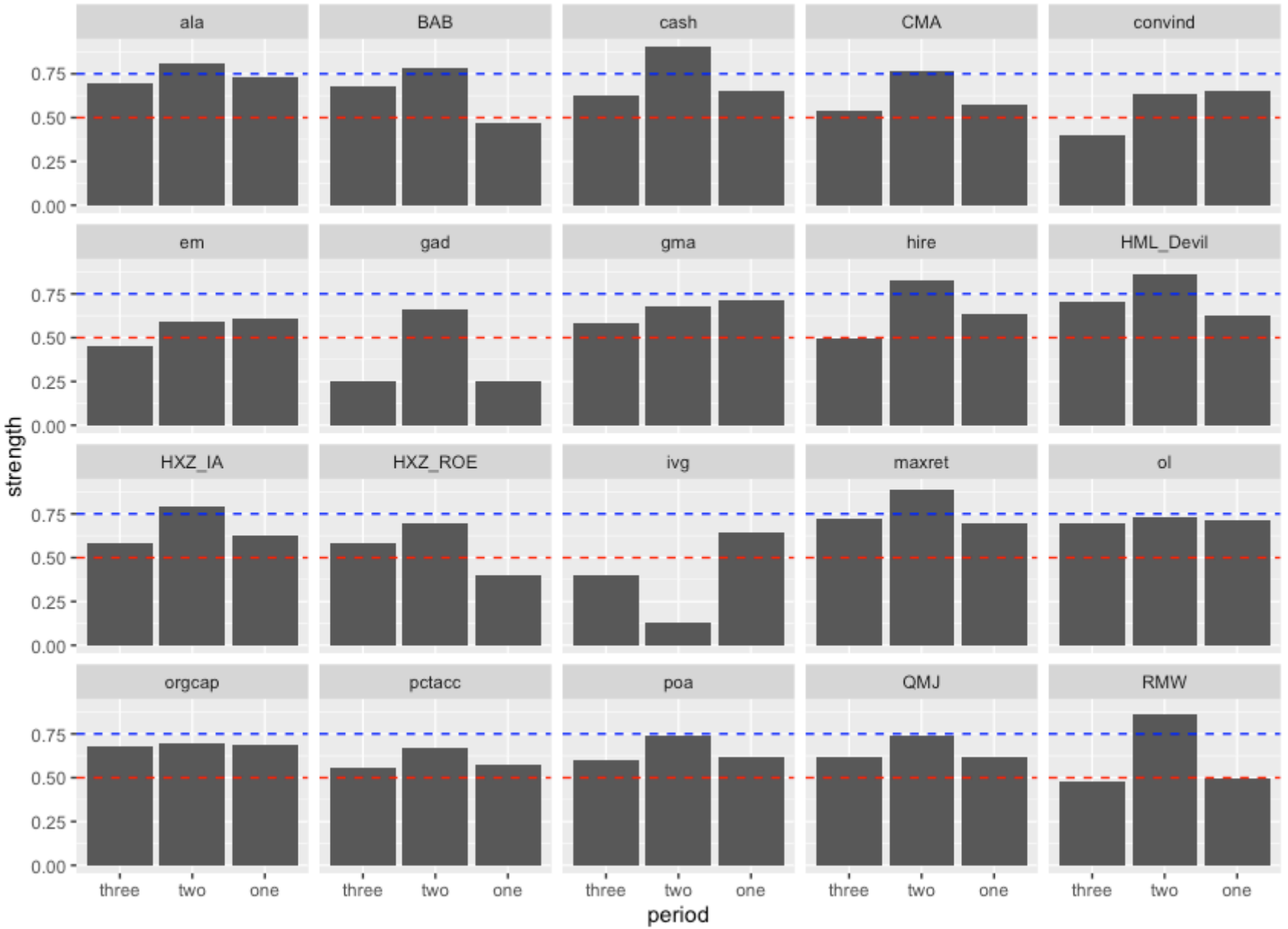



acc
bm_ia
cfp_ia
chempia
cp
cto
dolvol
etr
indmom
IPO
lfe
nincr
os
pchcapx_ia
pchgm_pchsale
pchsale_pchinvt
pchsale_pchrect
pchsale_pchxsga
ps
sp
std_dolvol
std_turn
turn
UMD
zs
0.00
0.25
0.50
0.75
strength
three
two
one
period


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
0.00
0.25
0.50
0.75
strength
three
two
one
period


dfnl
dwc
ndf
roic
zerotrade
dfin
dsti
ms
retvol
ww
dcol
dnco
herf
pchcapx3
tang
chmom
dncl
grcapx
nxf
ta
cei
dnca
egr_hxz
nef
rs
0.75
0.50
0.25
0.00
three
two
one
strength
period

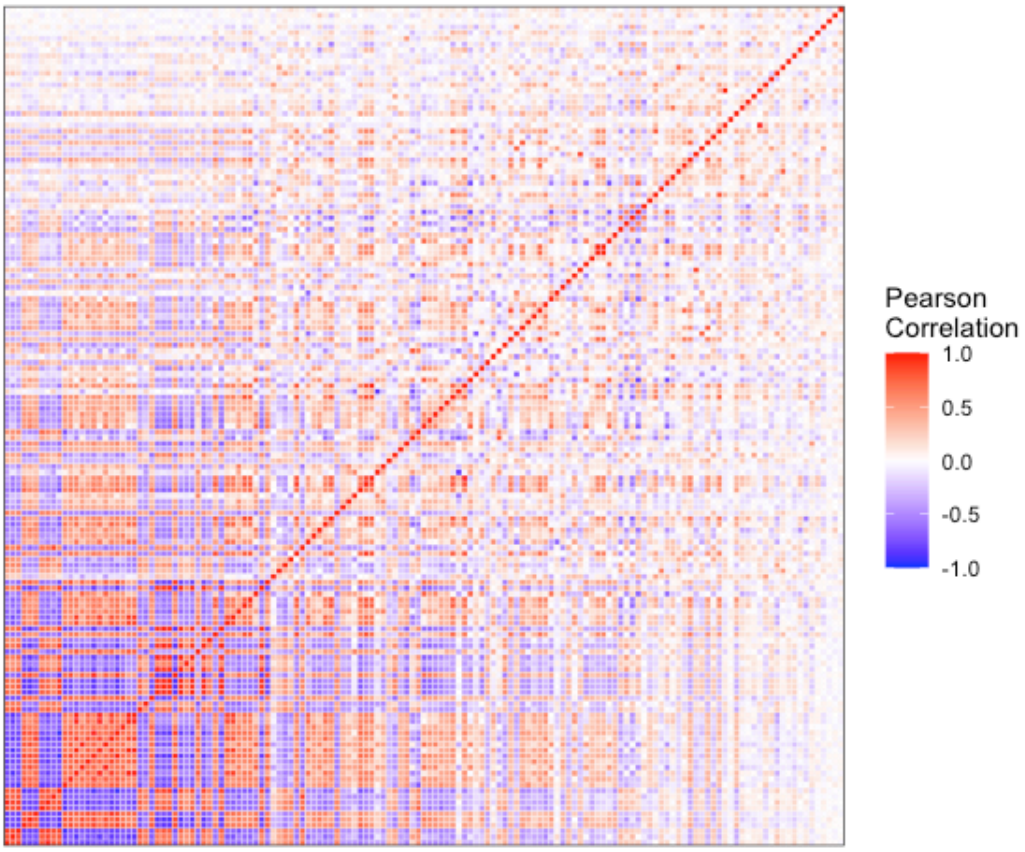

absacc
aeavol
ato
cashpr
cdi
chatoia
chcsho
chpmia
dpia
ear
ebp
invest
moms12m
ndp
nop
op
pchcapx
pm
ra
realestate_hxz
roaq
rsup
sin
stdacc
stdcf
0.00
0.25
0.50
0.75
strength
one
two
three
period



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.

B.3 Factor Correlation

Figure B.3: Risk Factors Correlation Coefficient



Notes: The figure visualize the correlation coefficient among 145 risk factors included in this paper. From the lower-left corner to the upper-right corner, the factor strength of each factors increases. We can see that the dark colours are clustering in the lower-left corner, this means that factor with strong strength present high correlation with other factor with strong strength. With the decrease of the factor strength, the correlation diminish.