A Generalized Machine Learning Framework for Linear Factor Model Test

Christopher Jones
University of Southern California
Email: christopher.jones@marshall.usc.edu

Jinchi Lv
University of Southern California
Email: jinchilv@marshall.usc.edu

Kuntara Pukthuanthong
University of Missouri
Email: pukthuanthongk@missouri.edu

Junbo Wang
Louisiana State University
Email: junbowang@lsu.edu

July 31, 2020

Abstract

We introduce a generalized statistical learning method, sparse orthogonal factor regression (SOFAR), in testing linear factor models with both large numbers of factors and testing assets. Our approach encompasses most of the existing methods in the literature and can be used in many other scenarios with large data sets. Applying SOFAR, we can select the latent factors from the whole swath of 219 candidate factors proposed by the literature simultaneously, identify test assets associated with the selected latent factors, and interpret them. We can also select the latent factors and correlated characteristics in the IPCA framework without bootstrapping. Without firm characteristics instrumenting, we find that four latent factors (market, investment, intangible, and frictions) are relevant to the covariance of asset returns and three types of factors (profitability, asset liquidity, and liquidity bets) price assets in cross-section. We also find that the out-of-sample prediction for the asset pricing model can be more precise with candidate factor selections. With characteristics as instruments, we only identify one factor, and the correlated characteristics are beta, size, momentum, and liquidity.

Keywords: Big data, principle components, factors, characteristics

JEL classification: G10, G11, G12

1 Introduction

Asset pricing literature has featured many anomalies or candidate risk factors. For example, Harvey, Liu, and Zhu (2016) identify more than 300 factors and recently has updated them to 440 (Harvey and Liu, 2019). Many of these candidate factors could proxy for a similar risk, leaving most of them redundant. In addition to the factors, the number of characteristics is also exploding (For example, Green, Hand, and Zhang (2017) examine more than 90 characteristics), the number of related covariance characteristics might also be much smaller than all characteristics identified in the literature.

To deal with the issue of redundancy in factors/characteristics, researchers typically apply two methods: one approach is to extract common information from all candidate factors or characteristics using PCA or single value decomposition method (e.g., Pukthuangthong, Roll and Subramanyam (2019), Stambaugh and Yuan (2017), Giglio and Xiu (2018), Kelly, Pruitt, and Su (2017, 2018), Huang, Li and Zhou (2019)). The other approach is to directly select risk factors from all candidate factors (e.g., Feng, Giglio, and Xiu (2019)). Given that all of these methods are intended to achieve similar goals, a natural question is whether there is a generalized framework that can encompass these methods but still retain the desired property of these methods.

Besides the number of the proposed factors, the number of testing assets has also been growing. Portfolios used for testing these factors are increasing dramatically. Chen, Roll, and Ross (1986) and many other researchers use ten size portfolios. Fama and French group the stocks into 49 industrial portfolios. They also create portfolios by double or triple sorting the stocks with different firm characteristics. Given that more and more firm characteristics with pricing effects are identified, the number of sorted portfolios may continue to grow. Also, a recent strand of literature advocates the use of individual stocks (Gagliardinia, Ossolab and Scaillet (2016), Jegadeesh, Noh, Pukthuanthong, Roll and Wang (2019), and Pukthuanthong, Roll, Wang and Zhang (2019)).

Although the number of individual assets grows exponentially, it is unclear whether these portfolios or individual stocks should be used to test all factor models. For example, the Fama French 25 portfolio sorted by size and book-to-market might not be useful to test factors that are uncorrelated with size or book-to-market. Including them in the test will only produce more noise in the estimation. In addition, Gospodinov, Robotti, and Kan (2018) find that the assets or portfolios that are uncorrelated with tested factors can lead to very significant risk premium estimation with close-to-1 R-square in a cross-sectional regression analysis. Therefore, it is also useful to select the correlated portfolios to test the candidate factors.

To generalize the methods, resolve redundancy of factors/characteristics, and deal with the uncorrelated portfolio issue, this paper introduces a new framework, sparse orthogonal factor regression (henceforth SOFAR) proposed by Uematsu, Fan, Chen, Lv and Lin (2019). The method can test models with a large number of factors and assets. SOFAR can control the number of principal components. Moreover, for each latent factor, SOFAR can sparsely select a few factors from all candidate factors, and the latent factor will be equal to a linear combination of these selected candidate factors. Also, SOFAR can select the testing assets that are correlated with each latent factor.

Since SOFAR can select variables in several dimensions, the method can be applied to a wide range of tests. For example, we can identify the number of important risk factors, i.e., select the non-redundant factors from a factor zoo, and pick the correlated test assets at the same time. In addition, we can also turn off some of the selection restrictions to resolve any of the single selection problem or any combinations of the selection problems above, making this approach flexible. The aforementioned methods proposed in factor testing are equivalent to applying one of the restrictions in SOFAR; hence, SOFAR encompasses these methods. Moreover, it can be applied to linear asset pricing tests that are not possible with existing methods when the data sets are large.

Specifically, we apply SOFAR to test various assets, pricing models. First, we can select the factors correlated with the risk of assets. Huang, Li, and Zhou (2019) introduce a reduced rank approach to select risk factors. They assume that the latent risk factors are linear combinations of the candidate factors. The latent factors are therefore determined by the largest eigenvalues of the coefficient of regressing asset returns on a large number of factor candidates. The method is similar to a simple singular value decomposition approach. However, they assume there are five latent factors without offering the selection guidelines. Moreover, each of the latent factors is a linear combination of all candidate factors. If there are 200 candidate factors, the linear combinations of them, which are latent factors, can be challenging to interpret. Finally, it is possible that some of the testing assets are not highly correlated with for latent factors; thus, it is hard to justify that they are risk-related.

SOFAR extends this approach with the selection criteria for the number of latent factors, the number of relevant candidate factors for each latent factor, and the testing assets. Hence, the method offers several benefits. First, we can select only important risk factors, given that the eigenvalues of the single value decomposition die out quickly over the latent factors. Second, it is possible to allow researchers to interpret each latent factor. For example, if most of the selected candidate factors for one latent factor are in the category of investment, the latent factor likely represents risk related to investment. In addition, we can select correlated testing assets for each latent factor at the same time. For example, if we assume that the risk factors should at least price more than 50% of the portfolios (different factors may be able to price different subsets of portfolios), SOFAR can select these factors. Huang, Li, and Zhou (2019) cannot achieve the same flexibility since they could not select the related candidate factors or stocks to each latent factor at the same time with the reduced rank method.

Giglio and Xiu (2018) propose a new method to estimate the risk premium of factors. The method is effective when there are missing factors. Specifically, they propose to obtain principal components of testing assets as latent risk factors. If a candidate factor generates risk premiums for assets, it should be written as a linear combination of the latent factors. Given that we could estimate the risk premium of all latent risk factors (which will not suffer missing factor issue), the estimated risk premium of the candidate factor is the same linear combination of the risk premium of the latent factors. When we have a set of preselected candidate factors that represent different risks, Gilgio and Xiu (2018) method is a perfect fit. However, given a large number of candidate factors, the method can give misleading results. For example, the investment factor and the value factor are highly correlated. When we apply Giglio and Xiu (2018), both factors can be projected to the latent factors similarly, and the estimated risk premiums, as well as the T-statistics for two factors, should also be quite similar. Therefore, it is possible to find that both investment and value factors generate similar significant risk premiums, although they might be essentially the same risk factor. Hence, their approach might not be able to identify the correct pricing factors when the factor zoo is large, and some of the factors are highly correlated.

The example above indicates that there is one missing step in Giglio and Xiu's method: factor identification when the number of candidate factors is large. The factor structure we obtain from SOFAR can naturally extend the Giglio and Xiu (2018) method. Specifically, for each latent factor, we can identify only a few candidate factors that are correlated with them. After we select these candidate factors, we can estimate their risk premium following the same procedures of Giglio and Xiu (2018). Hence, our method can identify factors and calculate the corresponding risk premium at the same time. Also, if the latent factors can be economically interpretable, we only need to estimate the risk premiums of the latent factors directly, and the economic theories can explain the factor structure from these latent factors.

Our empirical results demonstrate SOFAR advance the other existing approaches. To illustrate, we extend the reduced rank method in three dimensions: (1) we can select several latent factors, (2) with each latent factor, we can select correlated candidate factors, and (3) for each latent factor, we can select correlated testing assets. We find four to six latent factors representing the market, investment, intangible, and friction since the most candidate factors correlated with these latent factors are selected from these categories. Applying Giglio and Xiu (2018) to estimate the risk premium of these candidate factors, we find that only three types of true risk factors can price assets. They represent profitability, asset liquidity, and liquidity beta. We also test the out-of-sample prediction of the SOFAR methods in a horserace to the RRA method. When the candidate factor selection is included, the out-of-sample pricing error is smaller, and R-square is larger than the RRA method.

Another application of the SOFAR method is to extend the IPCA (Kelly, Pruitt, and Su (2017, 2018)) method. IPCA method can identify the factor structure for individual stocks, using characteristics as the instruments for time-varying factor loadings. However, there is no guideline on how many factors we should choose. Also, when selecting the characteristics that affect factor structure, their approach requires bootstrapping to generate corresponding statistical tests. This is because IPCA is a modified PCA method without controlling the number of PCs and sparsely selecting the correlated characteristics. SOFAR simplifies these steps, i.e., we can select the number of factors, and simultaneously identify the characteristics without the bootstrap method, which requires estimations of model coefficients. Since SOFAR assumes that the factor loading is linear on characteristics, we need to estimate the corresponding coefficient for each characteristic for each factor. When there are five factors and 90 characteristics, there are 450 coefficients to estimate. With data of less than 40 years, it is likely to have an over-identification issue. Hence, IPCA is more effective with a relatively small number of characteristics. SOFAR does not suffer the same issue, given that the model is inherently designed to deal with a large number of coefficients.

Empirically, we use individual stock returns and 90 characteristics following Green, Hand, and Zhang (2017). Extending the IPCA using SOFAR, we find that there is only one latent factor selected. This is consistent with Kelly, Pruitt, and Su (2019) since the in-sample R-square of IPCA does not increase significantly after the first latent factor. Also, the main correlated characteristics are beta, size, momentum, and liquidity, which is also similar to their results.

The contribution of the paper is two-fold. First, we introduce the SOFAR approach to the asset pricing test. SOFAR encompasses and extends the most current methods in the literature. Second, we provide empirical findings from applying SOFAR to a large set of candidate factors, and a large number of assets.

Classical asset pricing tests often deal with a small sample of testing assets (characteristics sorted portfolios or industry) and a small sample of factors (CAPM, Fama-French 5 to 6 factors, Q factors, or macroeconomic factors). With a surge of accounting and economic data, a large number of factors are discovered with economic theories (McLean and Pontiff (2016)). The development of these factors prompts a long-term argument in asset pricing study, which tries to determine the optimum factor structure. Based on the Sharpe ratio test, Barillas and Shanken (2017, 2018) develop the model comparison test, and Barillas, Kan, Robotti, and Shanken (2018), Fama and French (2018) and Ferson, Siegel and Wang (2019) further develop and apply this method. The idea is that the Sharpe ratio generated from the true factors should span the mean-variance efficient frontier of asset returns. Any other factors cannot contribute more to the optimum efficient portfolio. Another method is based on the cross-sectional analysis, where a priced factor should explain the cross-sectional stock returns. Since a large number of factors are discovered, factor selection becomes an important issue. This strand of studies is proliferating with various newly developed methods we describe in this paper. Our new approach can be viewed as the generalization of this line of researches. A third method is to test whether a stochastic discount factor constructed by economic variables or firm characteristics can price assets. For example, Kozak, Nagel, and Santosh (2018b) construct a stochastic discount factor from many stock characteristics, based on the sparsity assumption of priced characteristics.

Besides factor selection, machine learning methods (linear and non-linear) are applied in many other asset pricing types of research. Rapach, Strauss, and Zhou (2013) apply the Lasso approach to select the predictors of the international stock market. Chinco, Clark-Joseph, and Ye (2018) apply the Lasso approach to identify the short-lived signal to predict future stock returns from a sizeable intra-daily information set. Han, He, Rapach, and Zhou (2018) compare various variable selection methods to choose predictors of individual stocks from 94 firm characteristics. Rapach and Zhou (2019) use the sparse principal component approach to construct a macroeconomic factor structure. Given that these methods are based on linear prediction models, we expect that SOFAR can also be applied to generalize these methods. Based on a non-linear deep learning approach, Gu, Kelly, and Xiu (2019) compare various machine learning approaches for risk premium prediction. Moreover, Gu, Kelly, and Xiu (2019) develop a method to construct latent factors based on conditioning information. Although SOFAR is not directly applicable in a general non-linear model, if the non-linear prediction function can be approximately written as a polynomial function of the predictors or factors, SOFAR can be used in these scenarios as well[1].

2 Methodology

2.1 Introduction to SOFAR and generalization of Huang, Li, and Zhou (2019)

In this section, we introduce the Sparse Orthogonal Factor Regression (SOFAR) in the context of the asset pricing test. Let $\boldsymbol{R}$, a T times N matrix, be test asset, and $\boldsymbol{F}$, a T times M matrix, be candidate factors. Here, N is the number of assets, M is the number of candidate factors, and T is the number of periods. As we discussed in the introduction, both N and M are large. We assume that asset returns $\boldsymbol{R}$ are linear on candidate factors $\boldsymbol{F}$. Therefore, the testing model can be written as:

$$\boldsymbol{R} = \boldsymbol{F}\boldsymbol{B} + \boldsymbol{E}. \tag{1}$$

Here $\boldsymbol{B}$, the M times N matrix, is the regression coefficient matrix; and $\boldsymbol{E}$, the T times N matrix, is the regression residuals. Given that both N and M are large, the regression coefficient is a matrix

[1] More generally, if predictions can be non-parametric, we could apply SOFAR using B-spline (a linear combination of linear or non-linear functions) or wavelets (decompose the model into the frequency functions, following Fourier Analysis).

that contains a large number of elements. Note that this model is the same as Huang, Li, and Zhou (2019).

We assume that there are only a small number of true risk factors[2], even if there are many candidate factors. That is to say, in Equation (1), either there are only a small number of candidate factors or a small number of linear combinations of the candidate factors that will be related to returns. In the first scenario, most of the M candidate factors are not priced, so the rank of matrix $\boldsymbol{F}$ should be much smaller than min(M,T) (assume that the number of periods, T, is large enough). The second scenario is identified by Lowellen, Nagel, and Shanken (2011), specifically, the rotational indeterminacy of factors. They propose that the true factors can be linear combinations of candidate factors. Since the number of true risk factors is small, this implies that the rank of the coefficient matrix, $\boldsymbol{B}$, should be small.

To identify the true factors, we need to apply a variable selection or factor analysis (such as principal component) techniques. One classical method is Lasso, which can select a small number of factors with the highest correlation with asset returns. However, assume that most of the candidate factors are correlated with asset returns, and the true risk factors can be written as linear combinations of candidate factors (when there is rotational indeterminacy). In this case, applying Lasso will mistakenly remove most of the risk related candidate factors. To address this issue, Huang, Li, and Zhou (2019) apply a reduced-rank regression approach to identify these risk factors. However, a fraction of candidate factors may be redundant, and the true factors are still linear combinations of the remaining candidate factors. Moreover, the true factor might only be correlated with a fraction of assets, leaving a significant portion of the assets not affected by it. Huang, Li, and Zhou (2019) cannot resolve these issues.

SOFAR can handle all the aforementioned issues together. Specifically, instead of using penalty function to restrict the number of coefficients in the regression, SOFAR introduces three penalty functions to control the sparsity of factors (number of latent factors), the variable selection on returns (select correlated assets), and the variable selection on factors (select correlated candidate factors). More mathematically, assume that the regression coefficient matrix can be written as the following singular value decomposition. $\boldsymbol{B} = \boldsymbol{UDV'}$, where $\boldsymbol{D} = diag(d_1, \cdots, d_L)$ is an L by L

[2] It is arguable that the covariance of assets can represent the risk, as Kozak, S., Nagel, S., & Santoshi, S. (2018 1) suggests. Thus, the true risk factors in this paper are the factors that can represent the covariance matrix of the assets.

diagonal matrix, with $L = min\ (M, N)$. Assume $d_1 \geq d_2 \geq \cdots \geq d_L \geq 0$, and the number of the non-zero eigenvalues, K, is much smaller than L, given the sparsity of the true factors. Both $\boldsymbol{U}$ (M times L) and $\boldsymbol{V}$ (N times L) are orthonormal matrices (i.e.$\boldsymbol{U}^{'}\boldsymbol{U} = I$ and $\boldsymbol{V}'\boldsymbol{V} = I$, with $\boldsymbol{V}^{'}$ or $\boldsymbol{U}^{'}$ the transpose of matrix $\boldsymbol{V}$ or $\boldsymbol{U}$). For any matrix $\boldsymbol{M} = (m_{i,j})$, we define Frobenius norm as $\parallel \boldsymbol{M} \parallel_F = (\sum_{i,j} m_{i,j}^2)^{1/2}$ and L_1 norm as $\parallel \boldsymbol{M} \parallel_1 = \sum_{i,j} |m_{i,j}|$. Therefore, the orthogonality constrained optimization problem can be written as:

$$\left(\overset{\wedge}{\boldsymbol{D}}, \overset{\wedge}{\boldsymbol{U}}, \overset{\wedge}{\boldsymbol{V}}\right) = \underset{\boldsymbol{D},\boldsymbol{U},\boldsymbol{V}}{argmin} \left\{\frac{1}{2}(\parallel \boldsymbol{R} - \boldsymbol{FUDV'} \parallel_F + \lambda_d \parallel \boldsymbol{D} \parallel_1 + \lambda_a \rho_a(\boldsymbol{UD}) + \lambda_b \rho_b(\boldsymbol{VD}))\right\}$$

$$\text{subject to } \boldsymbol{U}^{'}\boldsymbol{U} = I \text{ and } \boldsymbol{V}'\boldsymbol{V} = I.^{3} \qquad (2)$$

Here, ρ_a and ρ_b are penalty functions. They are used to select the independent variables (candidate factors), and dependent variables (asset returns), respectively. Uematsu, Fan, Chen, Lv, and Lin (2019) show that they can be any convex functions. For example, if both functions are L_1 norm of the corresponding matrices, the selection criteria are similar to Lasso. If the functions are quadratic, the selection criteria are similar to Chinco, Neuhierl, and Weber (2020), which have a Bayesian interpretation. At the same time, $\parallel \boldsymbol{D} \parallel_1$ is used to select eigenvalues of the coefficient matrix. λ_d, λ_a and λ_b are tuning parameters that control the number of the selected variables.

Huang, Li, and Zhou (2019) (reduced-rank regression) is a special case of the setting above. Specifically, a simplified version[4] of their approach can be written as

$$\left(\overset{\wedge}{\boldsymbol{D}}, \overset{\wedge}{\boldsymbol{U}}, \overset{\wedge}{\boldsymbol{V}}\right) = \underset{\boldsymbol{D},\boldsymbol{U},\boldsymbol{V}}{argmin} \left\{\frac{1}{2}(\parallel \boldsymbol{R} - \boldsymbol{FUDV'} \parallel_F)\right\}. \qquad (3)$$

Note that in this case, we do not impose any constraints to coefficients, i.e., there is no latent factor selection, no candidate factor selection, and no testing asset selection. To estimate the parameters in equation (3), Huang, Li, and Zhou (2019) find the largest five eigenvalues of $\boldsymbol{BB'}$, which is equivalent to finding the largest five absolute eigenvalues of $\boldsymbol{B}$. Then, through singular value decomposition, $\boldsymbol{B} = \boldsymbol{UDV'}$, they can estimate the corresponding $\boldsymbol{U}$ and $\boldsymbol{V}$ matrices. These two

[3] Note that we present SOFAR in return factor model. The method is more general, and can be applied to any linear models, with **R** replaced by a dependent variable **Y**, and **F** replaced by an independent variable **X**.

[4] We simply the method by assuming the weighting matrix of the Principal Component Analysis to be identity and the conditioning information is not included.

matrices are used to determine the composition of each latent factor, and the linear combination of the testing asset returns correlated with the latent factors. Huang, Li, and Zhou (2019) assume that there are 5 latent factors (risk factors), without statistical guidelines. With the tuning parameter λ_d, SOFAR can select the number of principal components. Moreover, with parameter λ_a, we can select a fraction of candidate factors that are correlated to the risk factors. In case that some candidate factors are redundant or not risk-related, the method can be useful to identify them and exclude them from factor construction. This procedure will remove noise in the resulting factors. Moreover, it is useful for interpreting the results when the selected candidate factors are clustered in certain categories. With λ_b, we can select the testing assets that are correlated with the selected latent factors.

2.2 Complement of Giglio and Xiu (2018)

In the asset pricing test, we are not only interested in finding the factors that are correlated with assets, but also the factors that can price these assets in cross-section. A classical method we apply is the Fama-Macbeth approach. However, factors are likely to be correlated. If we do not include some factors (for example, these factors are not identified in academics yet) in the Fama-Macbeth regression, the estimated risk premium for the factors can be biased. Therefore, Giglio and Xiu (2018) propose a three-step method to estimate risk premiums that can address this omitted variable bias. Their method is as follows:

(1) Estimate the Principal Components (PCs) of the tested portfolio returns as latent factors

(2) Estimate risk premiums of these PCs

(3) For any test factor, project the factor into PCs. Then the risk premium of the factor is a linear combination of the risk premium of PCs.

The method can be applied for risk premium estimation of any factor. However, the asset pricing test also examines factor identification (Pukthuangthong, Roll, and Subramanyam (2019)). For example, if the Q-factor model (Hou, Xue, and Zhang (2019)) is true, the investment factor should explain the value effect, i.e., the value factor is a noisier version of investment factor. In this case, two factors are highly correlated. When we apply Giglio and Xiu (2018), both factors can be projected to the same latent factors, and their risk premiums can be calculated (step 3). Given the correlation between value and investment factors, the projection coefficients should be similar.

Since the risk premiums of the two factors are linear combinations of the projection coefficients and risk premiums of the latent factors, both factors likely have similar risk premium estimates. These results can be misleading in factor identification because if the investment factor is the true pricing factor, the value factor should have a zero risk premium based on the Q-factor model. Giglio and Xiu (2018) might not be able to identify the distinct factors when candidate factors are correlated because they do not control for other factors when calculating the risk premiums. Given that there are many candidate factors proposed by the literature, it is almost impossible to avoid this issue if we apply Giglio and Xiu (2018). For instance, when we apply Giglio and Xiu (2018) to all 200 candidate factors we constructed in section 6, we find more than 50 factors are priced. Some of these factors might be overlapped.

To resolve this issue, we can apply SOFAR. Explicitly, we assign **R** as the testing portfolio returns, **F** as the candidate factors. With SOFAR, we can identify orthogonal PCs from a whole swath of candidate factors at the same time. Our empirical works shown in Section 6 show the first PC is likely to be the market factor. In contrast, the other PCs are only related to a few candidate factors when we impose the candidate factor selection criterion. Given that only less than 30 candidate factors are correlated with these PCs, we can estimate the risk premium following Giglio and Xiu (2018) to find pricing factors. This can mitigate the issue of factor identification significantly because a noisy factor (such as value factor) will have a much smaller correlation with the PCs when the true risk factor (such as investment factor) is controlled (Lowellen, Nagel, and Shanken (2011)). Thus, the noisy factor is more likely to be excluded from the variable selection procedure embedded in SOFAR.

An alternative way to achieve the goal is to use PCA and variable selection method. i.e., estimate PCs, then use the variable selection method to determine the most related candidate factors (for example, Pelger and Xiong (2018)). This method could leave out risk factors. For example, assume that we select the five highest correlated candidate factors with the second PC. The remaining component can still represent a significant portion of the risk, compared with some candidate factors selected from the third or fourth PCs. This is because the second latent factor has significantly higher eigenvalue than the third factor. The remaining component of the second latent factor can represent a more critical risk than of the third latent factor. In this case, we could leave out the risk related factor in the model if the remaining part of the second factor is not included in other selected factors.

Applying SOFAR can be free of the issue above: (1) The constructed latent factors in SOFAR is the same as the linear combination of selected candidate factors, leaving no risk factor out in each latent factor. (2) Since these PCs are orthogonal, if we leave out some risk factors, we just need to add one more PC until we incorporate all the risk factors in the model. Therefore, we can avoid omitted variable bias because of the orthogonal latent factors.

2.3 Extension of IPCA

We discuss the IPCA model setting, the challenge for the PCA method, and the innovation of the IPCA method. Then we discuss the limitation of the IPCA method and our extension of the method.

2.3.1 IPCA model setting:

At each time t, the asset pricing model is characterized by:

$$\boldsymbol{R_t} = \boldsymbol{F_t}\boldsymbol{\beta_{t-1}} + \boldsymbol{\varepsilon}_t \tag{4}$$

Here, $\boldsymbol{R_t}$, a one by N vector, is the return of all assets at t; $\boldsymbol{F_t}$, a one by K vector, is the factors; $\boldsymbol{\beta_{t-1}}$, a K by N matrix, is the factor loadings; and $\boldsymbol{\varepsilon}_t$, a one by N vector, is the regression residuals. As before, we assume that there are T periods.

The returns of assets are available, but we do not have information on factors and factor loadings. The classical PCA method can construct both factors and beta. However, the assumption is that the loadings are constant. The key innovation in IPCA is to deal with the case that factor loading is time-varying, i.e.

$$\boldsymbol{\beta_{t-1}} = \boldsymbol{C}\boldsymbol{Z_{t-1}} \tag{5}$$

Here, $\boldsymbol{Z_{t-1}}$, an L by N matrix, is the firm characteristics; $\boldsymbol{C}$, a K by L matrix, is the constant coefficients. If the characteristics are known, IPCA can construct coefficients **C** and factors together.

2.3.2 Challenge for PCA method in this setting:

The objective function for the PCA method for identifying latent factors is:

$$\max_{\boldsymbol{C}}(NT)^{-1}\, tr\left(\sum_{t}(\boldsymbol{\beta}'_{\boldsymbol{t}-1}\boldsymbol{\beta}_{\boldsymbol{t}-1})^{-1}(\boldsymbol{\beta}'_{\boldsymbol{t}-1}\boldsymbol{R}_{\boldsymbol{t}}\boldsymbol{R}'_{\boldsymbol{t}}\boldsymbol{\beta}_{\boldsymbol{t}-1})\right)$$

Since factor loadings are time-varying, PCA cannot deal with this issue.

2.3.3 Innovation in IPCA:

Assume that $\frac{\boldsymbol{Z}_{t-1}\boldsymbol{Z}'_{t-1}}{N} = \boldsymbol{I}_L$, and define $\boldsymbol{X_t} = \boldsymbol{R_t}\boldsymbol{Z}'_{\boldsymbol{t}-1}/N$. Then the above objective function can be written as:

$$\max_{\boldsymbol{C}} T^{-1}\, tr\left(\sum_{t}(\boldsymbol{C}'\boldsymbol{C})^{-1}\boldsymbol{C}'(\boldsymbol{X}_{\boldsymbol{t}}\boldsymbol{X}'_{\boldsymbol{t}})\boldsymbol{C}\right)$$

With this new objective function, we can construct factor, coefficients **C** and factor loadings using PCA.

2.3.4 Limitation and extension of IPCA method:

There are at least two limitations to the IPCA method. First, the number of factors cannot be controlled. Second, the number of characteristics cannot be selected. In Kelly, Pruitt, and Su (2018), they examine the number of factors by examining the in-sample and out-of-sample R-square of the factor model. Moreover, they identify several characteristics through the bootstrap method.

To resolve these two issues, we can apply SOFAR to extend the IPCA model as follows.

For each t, define a one times L vector $\boldsymbol{Y_t} = (\boldsymbol{R_t}\boldsymbol{Z}'_{\boldsymbol{t}-1}) * (\boldsymbol{Z}_{\boldsymbol{t}-1}\boldsymbol{Z}'_{\boldsymbol{t}-1})^{-1}/N$. Let $\boldsymbol{Y} = [\boldsymbol{Y}_1; \boldsymbol{Y}_2; \cdots \boldsymbol{Y_T}]$. We can then rewrite the asset pricing model in equation (4) as

$$\boldsymbol{Y} = \boldsymbol{FC} + \boldsymbol{E} \tag{6}$$

Here **F** is a T by K matrix, and **C** is a K by L matrix. Both **F** and **C** are unknown. Our goal is to estimate the factors $\boldsymbol{F_t}$ and non-time varying coefficient matrix **C** simultaneously. This is a classical Sparse PCA problem, which is one of the applications of the SOFAR. Specifically, we apply SOFAR following Equation (8) of Uematsu, Fan, Chen, Lv, and Lin (2019).

$$\left(\hat{\boldsymbol{D}}, \hat{\boldsymbol{U}}, \hat{\boldsymbol{V}}\right) = \underset{\boldsymbol{D},\boldsymbol{U},\boldsymbol{V}}{argmin} \left\{\frac{1}{2}(\parallel \boldsymbol{Y} - \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}' \parallel_F + \lambda_d \parallel \boldsymbol{D} \parallel_1 + \lambda_b \rho_b(\boldsymbol{V}\boldsymbol{D}))\right\}$$

$$\text{subject to } \boldsymbol{U}'\boldsymbol{U} = I \text{ and } \boldsymbol{V}'\boldsymbol{V} = I. \tag{7}$$

The SOFAR approach can deal with the two issues in the IPCA method without examination of the R-square or bootstrap because we can select the latent factors by imposing λ_d, and select characteristics by imposing λ_b.[5]

Finally, another advantage of this approach is that there is no need to assume that $\frac{\boldsymbol{Z}_{t-1}'\boldsymbol{Z}_{t-1}}{N} = \boldsymbol{I}_L$, (only assuming nonsingularity for our formulation as long as L is smaller than N).

3 Method discussion

In this section, we discuss the advantages of the statistical properties of SOFAR over other existing methods.

3.1 SOFAR vs. Lasso

Lasso is the classical variable-selection method in statistics. In the asset pricing model, suppose that there are a large number of factors that can determine the asset returns. Lasso can be applied to select a small number of factors.

SOFAR is a generalized model setup to Lasso. Specifically, in equation (2), if we assume that (1) the number of assets (M) is one, (2) the constraints $\lambda_d \parallel \boldsymbol{D} \parallel_1 + \lambda_b \rho_b(\boldsymbol{V}\boldsymbol{D})$ are not imposed, and (3) the function $\rho_a(\boldsymbol{U}\boldsymbol{D})$ is the L_1 norm function, the SOFAR reduces to Lasso.

There are several advantages of SOFAR over Lasso. First, Lasso is an individual regression, with only one dependent variable (one asset). However, the information from the one asset has an impact on the other assets and should be captured in the model. Lasso cannot resolve this issue. SOFAR is a join regression (seemingly unrelated regression). Thus, the information between assets is incorporated in the regression. Second, Lasso can produce sparsity, i.e., select a sparse number of factors from a large pool, but does not provide a factor structure. SOFAR, on the other hand, can create the latent factor structure, by taking advantage of the multiple asset settings (we can

[5] As our discussion in section 2.1, ρ_b is used to select dependent variables.

only implement PC analysis in multiple asset scenario). Third, the tuning parameters using Lasso are based on cross-validation, which is known to have a high false-positive and false discovery rate. Hence, we might find over-selection of the factors using Lasso. This issue is resolved in the SOFAR. This is because the SOFAR method has a relatively flexible penalty, and a more structured regression model, which can make the parameter less sensitive; thus, reducing the over-selection issue of the Lasso approach.

3.2 SOFAR vs. PCA, Sparse PCA and PCA+Lasso

PCA is a widely applied approach for factor analysis in asset pricing. Specifically, we decompose the variance-covariance matrix of the asset returns into a series of components, ranked by the importance of them. Each component corresponds to a T time one vector (time-series vector), which describes the common risk of the assets. So a series of risk factors can be extracted from this method. SOFAR is also a generalized PCA method. In equation (2), if (1) F is an M-times-M identity matrix, and (2) all constraints are removed, it becomes PCA. Since the effect of each component on the variance-covariance matrix is degrading quickly, we can obtain a factor structure with the PCA method, with a small number of factors. However, latent factors are not interpretable without additional analysis. Also, we do not utilize the information from the factors, as we set F to be an identity matrix.

There are two expansions of the PCA method, which are intended to achieve the interpretability: sparse PCA and PCA+Lasso. The PCA+Lasso applies PCA to asset returns to construct latent factors. Then for each latent factor, we select a small number of the most correlated candidate factors.[6] Sparse PCA, on the other hand, constructs the latent factors from the candidate factors, and select the candidate factors in each latent factor simultaneously. The latent factors constructed by the sparse PCA method can be correlated. If we choose four latent factors following the sparse PCA method, these latent factors may represent a similar type of risk, which can lead to a missing risk factor issue. In addition, the sparse PCA method does not construct a latent factor from the asset returns, which leaves out the information from the asset returns. Because of this issue, the method also cannot select the correlated assets simultaneously. PCA+Lasso method suffers the

[6] This method is similar to Pelger and Xiong (2018).

same issue as we described in section 2.3, since the Lasso or variable selection method might not incorporate all important factors.

SOFAR can resolve all these issues. Specifically, SOFAR utilizes the information from the candidate factors, as well as the asset returns. Moreover, given that the latent factors constructed in SOFAR are exact linear combinations of the candidate factors, we do not leave out any crucial factors.

3.3 The robustness of the SOFAR results

One concern of the machine learning method is its robustness to the tuning parameter selections. SOFAR also applied an advanced technique to alleviate this issue. Specifically, the method applies the generalized information criteria (GIC) to select tuning parameters following Fan and Tang (2012). The paper shows theoretically (in the large sample) and through simulation (in the finite sample) that the GIC method can identify the correct model for fairly complicated penalty forms, while other information criteria, such as AIC or BIC, might not. Also, given that the SOFAR adopts flexible penalties (adapted weighted penalties), the method can achieve the actuate estimation. So the tuning parameters are less sensitive. In section 6, we find that the SOFAR results are quite robust to the tuning parameters.

4 Data

We apply different sets of data. First, our stock returns are from CRSP from 1962 to 2018. Similar to the standard procedure of screening stock returns in the existing literature, we collect monthly returns from the Center for Research in Security Prices (CRSP) and accounting information from the Compustat Annual and Quarterly Fundamental Files. We exclude financial firms and firms with negative book equity. We do not exclude stocks with prices per share lower than $1 or $5. That is, microcaps are included in our sample. Hou et al. (2019) do not exclude either, and they show most candidate factors are applied to microcaps. We apply the same screening criteria, and delisting returns similar to them. Second, we apply the returns of 202 portfolios Giglio and Xu (2019) use as our testing assets. They include 25 portfolios sorted by size and book-to-market ratio, 17 industry portfolios, 25 portfolios sorted by operating probability and investment, 25 portfolios sorted by size and variance, 35 portfolios sorted by size and net issuance, 25 portfolios sorted by size and accruals, 25 portfolios sorted by size and beta, and 25 portfolios sorted by size and momentum. We reason this set of portfolios captures a vast cross-section of candidate factors and exposures to different factors; at the same time, they are readily available on Dacheng's website, and therefore represent a natural starting point to illustrate our methodology. Third, our 219 candidate factors include the returns of long-short portfolios constructed based on candidate factors that are constructed similarly to those in Hou, Xue, and Zhang (2019). Their candidate factors cover six types, including momentum, value versus growth, investment, profitability, intangibles, and trading frictions. At this stage, we limit our candidate factors estimated from one month and they are value-weighted average. The long-short portfolios are created based on the decile portolios sorted by 219 characteristics. Thus, we can create 2190 decile portfolios. These portfolios are used for out-of-sample test purpose. Fourth, we use macro factors provided by Serena Ng and Jurado et al. (2015). Fifth, we collect characteristics and show their statistics in Table 1. Our 84 characteristics are the same as those in Green et al. (2017) as we apply the same code they use in their paper. Jeremiah Green makes it available on his website.[7] The statistics of candidate factors shown in Table 2 are close to those of Hou et al. (2019) as we follow their methodologies. The discrepancy might be attributed to the fact that we do not use Compustat/Merge as they do. We provide the descriptive statistics of individual stock returns and characteristics in Table 1.

5 Simulations

Coming soon!

6 Empirical Results

6.1 In sample results

The first application is to extend Huang, Li, and Zhou (2019). Following their paper, we use 202 portfolios from Giglio and Xiu (2019) as the testing asset. On the other hand, we construct 219 candidate factors following Hou, Xue, and Zhang (2019). We exclude 29 candidate factors with less than 90% of the data. (Given that SOFAR requires a full matrix of data, we need to set candidate factor value equal to zero if they do not exist. Therefore, if the time-series of the data

[7] https://drive.google.com/file/d/0BwwEXkCgXEdRQWZreUpKOHBXOUU/view

has too many missing values, we assign too many missing values to zero, and the results can be unreliable. Hence, we exclude these variables when there are too many missing values.)

We standardize these candidate factors before we apply SOFAR. Because the latent factor is a linear combination of candidate factors, the selection of the candidate factors is based on these linear coefficients. If we do not standardize factors, the high coefficient for a candidate factor might be the result of its low variance. By standardizing, the magnitude of the coefficient solely identifies the total impact of the candidate factor on the latent factor.

We conduct several different analyses, starting with the basis Huang, Li, and Zhou (2019) method. In this case, we assume that there are 5 latent factors similar to what they do. We do not select a sparse number of candidate factors that are related to each PC; thus, each PC can be a linear combination of all candidate factors. Also, we do not select testing assets that are correlated with each latent factor. The results of this simple case are shown in Table 3, Panel A. The first factor takes the majority of the cross-sectional covariance, as indicated by the eigenvalue of the first PC. This factor is quite close to a market return. We calculate the correlation between this factor and the market return and find that their correlation is 0.96. The second to fifth factors also have non-negligible eigenvalues. However, for all latent factors, all candidate factors are correlated with them. It is difficult to find which candidate factors are more important than others in determining the latent factor since there are factors in each category with high impact (coefficient) on latent factors. Hence, these latent factors cannot be interpreted economically.

*** Insert Table 3 Panel A here ***

The first extension is to control the number of PCs. We restrict λ_d in Equation (2). This restriction is similar to Bai (2003). We find that there are four to six factors selected based on a different value of tuning parameters. This is quite close to four or five factors implied by Q-theory (Hou, Xue and Zhang (2015), and Hou, Mo, Xue, and Zhang (2019)) or five or six factors empirical examined by recent researches (Barillas and Shanken (2018), Stambaugh and Yuan (2017), Barillas and Shanken (2018), Giglio and Xiu (2018), Huang, Li and Zhou (2019), and Fama and French (2018)).

The second extension is to impose the sparse selection of candidate factors in addition to factor restriction. Specifically, in Equation (2), we restrict λ_d and λ_a. We again find four to six factors, depending on tuning parameters. We present the result with five factors in Table 3, Panel B. Besides the eigenvalues of the latent factors, we also present the loadings (coefficients) of the candidate factors on each latent factor. I.e., each latent factor is a linear combination of the candidate factors with the coefficients shown in the Panel B. Given that we have standardized all the candidate factors before applying SOFAR, a higher loading indicates that the latent factor is higher correlated with the candidate factor. Each latent factor represents the candidate factors that have the highest coefficients.

Table 3 Panel B

The third extension is to include the selection of testing assets by imposing the restriction of λ_d, λ_a and λ_b (see Table 3 Panel C) i.e., we impose all three restrictions. If a latent factor represents the risk of the assets, it should be correlated with a large enough fraction of the testing assets. We do find this evidence. For the first factor, all testing assets are correlated with it even if we impose the asset selection constraint. For the rest factors, the numbers of correlated assets are 165, 135, and 125, respectively. If a latent factor is considered as the systematic factor, it is natural to assume that the factor should be correlated with a large enough number of the testing assets. The results above imply that more than 50% of the testing assets (in total, there are 202 assets) have large enough correlations with latent factors. Hence, all latent factors seem to represent the systematic component of testing assets. Besides, from Panel C, the number of latent factors and correlated candidate factors remain similar. For example, we find four latent factors and 27 correlated candidate factors.

Table 3 Panel C

From this panel, the most exciting finding is the economic interpretation of each latent factor. Except for the first latent factor, each of the other latent factors is a linear combination of a small number of candidate factors. For example, for factor 3, there are only nine candidate factors correlated with it. When the number of the correlated candidate factors is small, it is simpler to determine which economic factor influencing the latent factor, especially when the candidate factors selected into each latent factor have common economic meanings. For example, the largest coefficients of the second factor come from category 3 of Hou, Xue, and Zhang (2019) (anomaly 3.12, composite equity issuance, and 3.20, changes in book equity). The other two candidate factors with high coefficients are two specifications (net and return-return) of candidate factor

6.25, liquidity betas. Hence, both of them represent a very similar effect. Given that the two coefficients are opposite in sign with similar absolute values, the effect of one candidate factor is canceled by the other factor, and the total impact of these two candidate factors in the latent factor should be small. To further examine whether the two candidate factors are highly correlated, we calculate the correlation matrices of the all candidate factors selected into each latent factors, and they are shown in Table 4, Panel A, B and C. From the Panel A, the correlation between these two factors is 0.98. Besides candidate factors 3.12, 3.20, and 6.25, the rest of the factors have much smaller coefficients. Hence, the second latent factor is mainly determined by factors 3.12 and 3.20. Both of them are from the investment category following Hou, Xue, and Zhang (2019), the candidate factor likely represents the firm investment (A.3, see Table 2).

*** Insert Table 4 here ***

Following the same logic, the third factor is mainly represented by candidate factors in category 5 (A.5, Intangibles). Hence, it can be interpreted as the intangibles effect. The fourth factor is mainly determined by the factors in both categories 5 and 6 (A.5 Intangibles and A.6 Trading frictions); therefore, there is evidence of the effect of the trading frictions on asset returns.

To further examine whether the latent factors represent a similar economic effect, we estimate correlation coefficients of the candidate factors selected into the latent factors and present them in Table 4. For all three latent factors, we can find that the correlations between the candidate factors with coefficients of the same (opposite) signs are mainly positive (negative). Therefore, if the latent factor represents the risk of the stock returns, the risks are mainly aligned with the risks associated with the candidate factors.

With these PC and candidate factors, we can further estimate the risk premium. When the latent factor has an economic interpretation, we can estimate their risk premium directly using Fama-Macbeth regression. The results are shown in Panel A of Table 5. We find that for the first three latent factors, although they are the factors representing a majority part of the cross-sectional covariance of returns, their risk premiums are insignificant. The fourth factor does represent a significant effect in pricing cross-sectional returns. Giglio and Xiu (2018) argue that there can be missing-variable bias in risk-premium estimations, and they propose a 3-pass method to estimate the risk premium. By applying their approach, the first and the second factors are still insignificantly priced, while the third and fourth factors are significantly priced. To further explore this result, we examine the loadings/betas of all testing assets for each latent factor. For the first two factors, the average values of the factor loadings are -0.0626 and -0.0019, while the standard deviations are 0.0089 and 0.0109. For the third and fourth factors, the average values are 0.0001 and 0.0003, but the variances are 0.0034 and 0.0043. Therefore, the loadings for the third and fourth factors are much more dispersed around their means, compared with that for the first two factors. Hence, the first and the second latent factors, which consist of a large portion of covariance of returns are likely to have similar loadings (This result is consistent with Keloharju, Linnainmaa, and Nyberg (2020), as they find that over the long term, the assets exhibit similar risk (variance)). When testing assets have similar loadings, the risk premium estimation can be subjected to estimation errors (Shanken (1992), and Pukthuanthong, Roll, Wang, and Zhang (2019)), lowering the power of the pricing test.

***Insert Table 5 Panel A ***

In addition, we can estimate the risk premium of candidate factors applying Giglio and Xiu (2018). We present the estimated risk premium following the three-pass method in the Panel B. Specifically, there are only a total of 27 distinct candidate factors (from 178) that are correlated with asset returns, following Panel C of Table 3. We find that there are only seven candidate factors with significant risk premiums. Most of them are from category 5 (intangibles), and category 6 (friction), which are the major components of factors three and four. A detailed examination of these individual factors shows more economic intuition of the factor structures. The first sets of the factors that can price assets in cross-section are operation leverage (5.8) and momentum (5.51). Recent literature shows that these factors are related to firm productivity (Novy-Marx (2011), Hou, Xue, and Zhang (2015), and Kogan, Li and Zhang (2020)). The second set of pricing factors are financial constraints (5.30) and asset liquidity (5.46). The third set of pricing factors are liquidity betas (6.25), which is related to the sensitivity of the firm return to the aggregate liquidity. Hence with the SOFAR and Giglio and Xiu (2018), we can roughly identify a three pricing-factor structure.

The results in Tables 3 and 5 can be viewed as an extension of the screening of price and risk factors by Pukthuangthong, Roll, and Subramanyam (2018). Specifically, the intangibles and friction factors seem to be correlated with the risk of the testing assets and can price their cross-sectional returns.

One issue of the machine learning method can be tuning parameter dependence. Therefore, we examine whether the results are robust to different tuning parameters. There are three variables to control the tuning parameters λ_d, λ_a and λ_a: the maximum total number of latent factors, and the upper and lower bound for the three tuning parameters. The maximum total number of latent factors provides an upper bound for the total number of latent factors. For example, we can set it equal to 10, which implies that we cannot select more than 10 latent factors, but we might only select 5 latent factors by imposing λ_d,. The maximum total number of latent factors guarantee that for any λ_d, the number of PC is less than or equal to 10. [8] The upper and lower bound for the three tuning parameters are used to controlling the tuning parameter directly. The SOFAR automatically estimates the maximum values of the three tuning parameters. The upper and lower bounds control the percentage of the maximum parameters. For example, if the upper bound is 0.95 and the lower bound is 0.05, the tuning parameter should be between 0.05 and 0.95 of the maximum values of these parameters.

We first examine the dependence of the results to the maximum total number of latent factors. In Panel A of Table 6, we set it to 8. The results are similar to those of 10, as we set in Table 3. We also try a few other values, and the results are quite similar. In addition, the upper and lower bound for the three controlling tuning parameters do not significantly affect the results, when they are larger than 0.95 or lower than 0.05, as the default value. In Panel B, we set them to be 0.97 and 0.03, and find that we can select five factors, and the four latent factors beyond the first factor have a similar selection of candidate factors. For example, the candidate factors with the highest coefficients for the second to fourth factors are almost the same. The candidate factors in each of the latent factors might not come from the same category, but they are still mainly coming from categories 3, 5, and 6. Hence, the latent factors can be the rotational factors from factors representing investment, intangibles, and frictions. Given that the controlling parameters are optimized internally in SOFAR, it is not advisable to set these two values too different from the optimized default values. Hence, values that are close to 0.95 and 0.05 should be a good upper and lower bound.

***Table 6 Panel A ***

[8] This is different from the total number of selected latent factors.

Next, apply SOFAR to study the factor model with time-varying loadings. We use the individual stock return together with 94 characteristics from Green, Hand, and Zhang (2017). We exclude ten characteristics with less than 90% of the data following the same logic in candidate factor selection; thus, we are left with 84 characteristics. Kelly, Pruitt, and Su (2019) use 36 characteristics. Their characteristics selection is based on model simulations, so they need to estimate the coefficients of the characteristics. When there is five-factor with 36 characteristics, the number of coefficients to be estimated in the model is five times 36 (which is 180). There can be an overfitting issue if we have more characteristics when the total sample period is less than 600 months. We can expand their sample of characteristics significantly because our method inherently select characteristics and simultaneously estimate the model, and does not depend on the bootstrap. Following Equation (7) in Section 2, we select both latent factors and characteristics. The results for different samples are presented in Table 7. First, we find that there is only one latent factor, when the coefficients of the factor are time-varying, for almost all subsamples (1980-2018, 1970-2015, 1995-2018, and 1980-2003). We also examine different tuning parameters and find that the number of factors is always equal to one. This result is consistent with Table 2 of Kelly, Pruitt, and Su (2019). They find that the in-sample R-square does not increase significantly after the first factor. SOFAR method creates selection criteria for factors with reasonable in-sample R-square. If the in-sample R-square is not significantly affected by additional factors, SOFAR will not select these factors. This result is also consistent with Huang, Li, and Zhou (2019). They also show that the factors beyond the first do not contribute to the factor structure significantly when the characteristics are included in time-varying factor loadings. Economically, this result indicates that the characteristics can be the main determinants of the time-varying expected return. Also, SOFAR can select 6-11 characteristics based on different tuning parameters. The characteristics being consistently selected from all subsamples are beta, size, momentum, and liquidity. These are also the major characteristics selected by Kelly, Pruitt, and Su (2019). Given that our method does not require the bootstrap method, we can significantly simplify the selection procedure, and we can select from a much larger number of characteristics.

***Table 7 here ***

6.2 Out-of-sample prediction

We also examine the out-of-sample prediction results by applying SOFAR. The application is to compare the explanatory power of various extensions of the RRA method.

To horserace of RRA and its extensions contain the following methods: the RRA method with constraints on latent factors, asset, and candidate factor selection, the RRA together with selecting a pre-specified the number of latent factors (five factors in this case), the RRA with a constraint on latent factors, the RRA with constraints on latent factors and asset selection, and the RRA with constraints on latent factors and candidate factor selection. These methods are associated with equation (2) with all λ_d, λ_a and λ_b turned on, equation (3) with the first five latent factors, with equation (2) with only λ_d turned on, with equation (2) with λ_d and λ_b turned on, and with equation (2) with λ_d and λ_a turned on. Following Huang, Li, and Zhou (2019), we choose the first 30 years as the training periods and remaining years for out-of-sample prediction. We also present the sum of alpha squares and the total R-square over all 202 portfolios, like Huang, Li, and Zhou (2019) did.

The results are shown in Panel A of Table 8. The RRA with the first five factors has a sum of alpha squares equal to 0.28 and an R-square of 0.55. The method leads to a better prediction than the RRA with a constraint on latent factors (sum of alpha squares 0.36 and R-square 0.31). This implies that the five-factor assumption for the prediction period seems to be helpful for prediction compared with the selection of the factors in the training sample (with the data from the first 30 years). However, if we impose the additional selection for the candidate factors, both the sum of alpha squares and R-square can be improved. Specifically, the RRA method with constraints on latent factors, asset, and candidate factor selection has the sum of alpha squares 0.10 and R-square 0.58, and the RRA with constraints on latent factors and candidate factor selection has the sum of alpha squares 0.07 and R-square 0.59. These results are also reasonable, because including too many candidate factors in latent factors likely generates more noise, which can negatively affect the prediction power.

We also apply the factors to 2190 portfolios sorted by 219 characteristics from Hou, Xue and Zhang (2019). The results are almost the same. We do find that the sum of alpha squares and R-square can be improved using SOFAR with candidate factor selections.

***Table 8 here ***

7 Conclusion

Current literature features a new and large body of asset pricing model tests with large data sets (with a large number of testing assets and candidate factors). These methods are mainly based on two approaches: 1 construct the latent factors that represent the risk of a large number of assets or the common component of a large number of factors. 2 select a few numbers of factors or characteristics using variable selection method or bootstrap method.

In this paper, we introduce a general framework for the test of the linear asset-pricing model, when there are a large number of test assets and a large number of candidate factors. The framework can combine the two methods mentioned above simultaneously. Therefore, it can encompass and expand these methods.

We show examples of applying this framework. Expanding Huang, Li, and Zhou (2019), we can study the economic interpretation of the latent factors. And we can find whether these factors can be correlated with a large number of testing assets. We can even select three categories of factors that can price assets. In out of sample prediction, we find that including the candidate factor selection is important in explaining the asset returns. Expanding the IPCA method, we find that there is only one factor when betas are time-varying and depending on the firm characteristics, which is consistent with a few studies in the recent literature. We can also select characteristics without the bootstrap method, from a large number of characteristics.

References

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica* 71(1), 135-171.

Barillas, F., Kan, R., Robotti, C., & Shanken, J. A. (2017). Model comparison with Sharpe ratios. Working paper.

Barillas, F., & Shanken, J. (2017). Which Alpha? *Review of Financial Studies 30,* 1316-1338.

Barillas, F., & Shanken, J. (2018). Comparing asset pricing models. *Journal of Finance*, *73*(2), 715-754.

Chen, N. F., Roll, R., & Ross, S. A. (1986). Economic forces and the stock market. *Journal of Business*, 383-403.

Chinco, A., Clark-Joseph, A., & Ye, M. (2018). Sparse Signals in the Cross-Section of Returns. *Journal of Finance, 74(1), 449-492.*

Chinco, A., Neuhierl, A., & Weber, M. (2020). Estimating The Anomaly Baserate. Working paper.

Fama, E. F., & French, K. R. (2018). Choosing factors. *Journal of Financial Economics*, *128*(2), 234-252.

Fan, Y., & Tang, C. (2012). Tuning parameter selection in high dimensional penalized likelihood. Journal of the Royal Statistical Society Series B 75, 531-552.

Feng, G., Giglio, S,. & Xiu, D. (2019). Taming the Factor Zoo: A Test of New Factors. Forthcoming *Journal of Finance.*

Ferson, W., Siegel, A., & Wang, J. (2019). Asymptotic Variances for Tests of Portfolio Efficiency and Factor Model Comparisons with Conditioning Information. Working paper.

Gagliardini, P., Ossola, E., & and Scaillet, O. (2016). Time-varying risk premium in large cross-sectional equity datasets. *Econometrica* 84(3): 985–1046.

Gagliardinia, P., Ossolab, E., & Scaillet, O. (2017). A Diagnostic Criterion for Approximate Factor Structure. Working paper.

Giglio, S., & Xiu, D. (2018). Asset Pricing with Omitted Factors. Working paper.

Gospodinov, N., Kan, R., & Robotti, C. (2018). Too good to be true? Fallacies in evaluating risk factor models. *Journal of Financial Economics.* 132(2), 451-471.

Green, J., Hand J., & Zhang, F. (2017). The Characteristics that Provide Independent Information about Average U.S. Monthly Stock Returns. *Review of Financial Studies.* 30(12), 4389-4436.

Gu, S., Kelly, B., & Xiu, D. (2019). Empirical Asset Pricing via Machine Learning. Forthcoming *Review of Financial Studies*.

Gu, S., Kelly, B., & Xiu, D. (2019). Autoencoder Asset Pricing Models. Forthcoming *Journal of Econometrics*.

Han, Y., He, A,. Rapach, D., & Zhou, G. (2018) What Firm Characteristics Drive US Stock Returns? Working paper.

Harvey, C., Liu, Y., & Zhu, H. (2016). …and the Cross-Section of Expected Returns. *Review of Financial Studies, 29(1), 5-68*.

Harvey, C., Liu, Y. (2019). A census of the factor zoo. Working paper.

Hou, K., Mo, H., Xue, C., & Zhang, L. (2019). Which Factors? *Review of Finance, 1-35*.

Hou, K., Xue, C., & Zhang, L. (2019). Replicating Anomalies. *Review of Financial Studies*. *Forthcoming.*

Huang, D., Li, J., & Zhou, G. (2019). Shrinking Factor Dimension: A Reduced-Rank Approach. Working paper.

Jegadeesh, N., Noh, J., Pukthuanthong, K., Roll, R., & Wang, J. (2019). Empirical tests of asset pricing models with individual assets: Resolving the errors-in-variables bias in risk premium estimation. *Journal of Financial Economics*, 133(2), 273-298.

Kelly, B., Pruitt, S., & Su, Y. (2017). Instrumented Principal Component Analysis. Working paper.

Kelly, B., Pruitt, S., & Su, Y. (2018). Characteristics are Covariances: A Unified Model of Risk and Return. *Journal of Financial Economics, 134(3), 501-524*.

Keloharju, M., Linnainmaa J., & Nyberg, P. (2020). Long-term discount rates do not vary across firms. *Working paper*.

Kogan, L., Li, J., and Zhang, H. (2020). Operating hedge and gross profitability premium. *Working Paper*.

Kozak, S., Nagel, S., & Santoshi, S. (2018 1) Interpreting Factor Models. *Journal of Finance, 73(3), 1183-1223*.

Kozak, S., Nagel, S., & Santoshi, S. (2018b) Shrinking the Cross Section. *Journal of Financial Economics*. *Forthcoming.*

Lettau, M., & Pelger, M. (2018). Estimating Latent Asset Pricing Factors. Working paper.

Lewellen, J., Nagel S., & Shanken, J. (2010). A skeptical appraisal of asset pricing tests. *Journal of Financial Economics*, 96, 175-194.

MacKinlay, Craig. (1995). Multifactor models do not explain deviations from the CAPM. *Journal of Financial Economics 38, 3-28*.

McLean, D., & Pontiff, J. (2016). Does Academic Research Destroy Stock Return Predictability? *Journal of Finance, 71(1), 5-32.*

Novy-Marx (2011). Operating leverage. *Review of Finance* 15(1), 2011, 103-134.

Pelger, M, & R, Xiong (2018). Interpretable Sparse Proximate Factors for Large Dimensions. *Working paper.*

Pukthuanthong, K., Roll, R., Wang, J., & Zhang T. (2019). A Tool Kit for Factor-Mimicking Portfolios. *Working paper.*

Pukthuanthong, K., Roll, R., & Subrahmanyam (2019). A Protocol for Factor Identification. *Review of Financial Studies, 32(4), 1573-1607.*

Rapach, S., Strauss, J., & Zhou, G. (2013). International Stock Return Predictability: What is the Role of the United States? *Journal of Finance, 68, 2013, 1633--1662.*

Rapach, S., & Zhou, G. (2019) Sparse Macro Factors. Working paper.

Shanken, J. (1992). On the estimation of beta-pricing models. *Review of Financial Studies* 5(1), 1-55.

Stambaugh, R., & Yuan, Y. (2017). Mispricing Factors. *Review of Financial Studies,* 30(4), 1270-1315.

Uematsu, Y., Fan, Y., Chen, K., Lv, J. and Lin, W. (2019). SOFAR: large-scale association network learning. *IEEE Transactions on Information Theory*, to appear.

Table 1: Summary statistics for portfolio returns and stock characteristics

This table presents descriptive statistics of the individual stock returns, and characteristics we apply in this study. Our sample period is from 1962 to 2018. Panel A presents the statistics of firm stock returns, including year, number of observations (# obs), number of stocks (# stocks), mean median, standard deviation (stdev), skewness, kurtosis, min and max of returns. Panel B presents the statistics of firm-level characteristics generated from Jeremiah Green's code. The statistics include characteristic name, mean, median, min, max, skewness, and kurtosis.

Panel A: Firm statistics

year	# obs	# stocks	mean	median	stdev	Skewness	kurtosis	min	max
1962	5	1	-0.003	-0.044	0.077	1.321	0.786	-0.060	0.120
1963	1741	334	0.007	0.000	0.065	0.864	2.848	-0.201	0.380
1964	3433	413	0.017	0.011	0.066	2.672	34.045	-0.361	1.106
1965	3327	313	0.022	0.015	0.072	1.472	8.720	-0.224	0.787
1966	4289	426	-0.004	-0.012	0.084	0.902	3.736	-0.410	0.602
1967	8284	1022	0.036	0.017	0.112	1.755	8.154	-0.317	1.143
1968	13413	1184	0.029	0.015	0.113	1.401	6.226	-0.368	1.284
1969	15023	1355	-0.021	-0.026	0.108	0.711	2.996	-0.573	0.866
1970	17221	1545	-0.003	-0.006	0.137	0.555	2.738	-1	1.240
1971	19514	1737	0.017	0.002	0.121	1.121	4.413	-0.887	1.000
1972	20997	1849	0.005	-0.004	0.110	1.087	7.244	-1	1.222
1973	25139	2714	-0.032	-0.037	0.141	0.567	3.064	-1	1.265
1974	35396	3216	-0.022	-0.035	0.154	1.834	16.816	-1	3.167
1975	37163	3230	0.049	0.013	0.178	2.358	15.083	-1	3.600
1976	36616	3224	0.037	0.012	0.142	1.994	12.297	-1	2.455
1977	36937	3208	0.016	0.000	0.118	4.332	95.630	-1	4.600
1978	36003	3165	0.021	0.012	0.143	1.510	17.649	-1	3.185
1979	36169	3296	0.030	0.013	0.131	1.974	15.909	-1	2.364
1980	38540	3530	0.031	0.015	0.157	2.209	22.061	-1	2.923
1981	39934	3596	0.004	0.000	0.136	1.905	19.609	-1	2.750
1982	41697	3836	0.023	0.000	0.163	2.884	35.633	-0.937	3.955
1983	43461	3896	0.029	0.010	0.170	3.469	40.525	-1	4.000
1984	44718	4208	-0.008	-0.005	0.143	1.922	29.865	-1	3.267
1985	46889	4304	0.019	0.003	0.218	66.773	9546.964	-1	31.764
1986	46475	4274	0.005	0.000	0.172	7.145	312.761	-1	10.200
1987	46944	4388	-0.003	0.000	0.193	3.215	89.871	-1	8.071
1988	49028	4583	0.017	0.000	0.164	4.423	98.139	-1	6.167
1989	48966	4456	0.011	0.000	0.161	4.055	88.530	-1	6.385
1990	48058	4353	-0.017	-0.016	0.195	7.449	288.474	-1	11.000
1991	47597	4296	0.038	0.010	0.230	10.021	394.427	-1	14.000
1992	47518	4349	0.021	0.000	0.205	7.289	197.868	-1	10.000
1993	49704	4527	0.018	0.000	0.170	5.377	133.287	-1	7.480
1994	55815	5448	-0.002	-0.005	0.154	11.954	816.860	-1	12.500
1995	62743	5794	0.024	0.013	0.160	3.339	51.066	-0.992	4.667
1996	64675	5977	0.016	0.005	0.166	3.994	93.064	-0.944	7.000
1997	67007	6295	0.020	0.010	0.173	3.560	80.276	-0.990	6.077
1998	67413	6273	0.002	-0.006	0.223	9.316	381.403	-0.974	12.667
1999	64586	6006	0.024	-0.004	0.232	5.167	96.251	-0.984	9.500

2000	62001	5841	-0.001	-0.011	0.256	3.848	72.772	-1	10.344
2001	59706	5590	0.026	0.006	0.266	3.921	54.947	-1	8.667
2002	56636	5053	-0.008	-0.009	0.212	2.669	36.205	-1	5.640
2003	52840	4710	0.048	0.024	0.186	3.745	49.567	-1	5.179
2004	50300	4463	0.018	0.010	0.149	5.980	175.177	-1	6.908
2005	49277	4412	0.004	0.000	0.131	2.075	31.445	-1	3.303
2006	48744	4364	0.014	0.007	0.128	2.687	45.132	-0.947	4.023
2007	47751	4318	-0.004	-0.005	0.131	3.825	124.273	-0.981	5.801
2008	46884	4232	-0.044	-0.035	0.199	1.490	24.403	-1	4.900
2009	45551	4013	0.046	0.020	0.274	10.771	427.327	-0.998	15.774
2010	43574	3829	0.023	0.015	0.157	3.312	72.655	-0.965	6.107
2011	41992	3714	-0.005	-0.008	0.147	2.690	44.459	-1	3.994
2012	40797	3587	0.016	0.010	0.145	4.732	103.288	-0.957	5.210
2013	39728	3491	0.032	0.023	0.131	3.586	59.621	-0.936	3.746
2014	39588	3510	0.005	0.004	0.135	6.295	192.536	-0.966	6.080
2015	40390	3643	-0.005	-0.005	0.154	11.459	609.285	-0.954	9.564
2016	40523	3602	0.016	0.010	0.170	8.124	273.583	-0.953	7.635
2017	39808	3509	0.013	0.007	0.142	3.251	47.289	-0.962	3.273
2018	39346	3483	-0.011	-0.008	0.154	4.699	129.833	-0.994	6.428

Panel B: Firm characteristics

Characteristic Name	Description	min	max	mean	median	skewness	kurtosis
acc	Working capital accruals	-1.02	0.58	-0.02	-0.02	-0.88	4.81
aeavol	Abnormal earnings announcement volume	-1.00	21.69	0.87	0.30	3.64	18.51
age	# years since first Compustat coverage	1.00	56.00	12.72	9.00	1.36	1.55
agr	Asset growth	-0.68	5.85	0.15	0.08	4.26	30.02
baspread	Bid-ask spread	0.00	0.91	0.05	0.03	5.15	38.11
beta	Beta	-0.74	3.94	1.08	1.01	0.69	0.69
bm	Book-to-market	-2.35	7.81	0.77	0.60	2.48	11.46
cash	Cash holdings	0.00	0.98	0.16	0.07	1.89	3.15
cashdebt	Cash flow to debt	-7.71	2.23	0.07	0.13	-4.14	25.99
cashpr	Cash productivity	-520.62	600.28	-1.90	-0.73	0.89	29.15
cfp	Cash flow to price ratio	-513.56	156.76	0.05	0.05	-172.20	57390.53
cfp_ia	Industry-adjusted cash flow to price ratio	-449.37	7031.61	13.09	0.00	21.92	479.82
chadv		-1.59	2.02	0.05	0.03	0.50	8.14
chatoia	Industry-adjusted change in asset turnover	-1.43	1.19	0.00	0.00	-0.15	4.74
chcsho	Chane in shares outstanding	-0.89	2.57	0.11	0.01	3.28	13.85
chfeps	Change in forecasted EPS	-6.48	8.25	0.00	0.00	1.29	121.37
chinv	Change in inventory	-0.29	0.37	0.01	0.00	1.10	6.77
chnanalyst	Change in number of analysts	-12.00	9.00	-0.01	0.00	-0.60	9.46
chtx	Change in tax expense	-0.12	0.16	0.00	0.00	0.35	13.06
cinvest	Corporate investment	-26.83	27.87	-0.02	0.00	-2.17	244.24
currat	Current ratio	0.16	60.34	3.16	2.00	5.58	40.35
depr	Depreciation/PP&E	0.01	5.51	0.26	0.15	5.92	49.70
disp	Dispersion in forecasted EPS	0.00	10.00	0.15	0.04	6.48	58.18
dy	Dividend to price	0.00	0.35	0.02	0.00	2.67	10.86
ear	Earnings announcement return	-0.46	0.51	0.00	0.00	0.26	3.17
egr	Growth in common shareholder equity	-3.54	8.19	0.14	0.08	3.32	28.51
ep	Earnings to price	-7.66	0.68	-0.01	0.05	-8.11	107.23
fgr5yr	Forecasted growth in 5-year EPS	-43.50	99.41	16.35	14.50	1.50	5.47
gma	Gross profitability	-0.84	1.78	0.37	0.33	0.81	1.52
grcapx	Growth in capital expenditures	-13.89	55.54	0.89	0.14	5.60	45.95
grGW		-0.92	11.69	0.12	0.00	8.06	83.62

grltnoa	Growth in long term net operating assets	-0.61	1.18	0.09	0.06	1.64	7.48
herf	Industry sales concentration	0.01	1.00	0.08	0.05	3.10	11.91
hire	Employee growth rate	-0.74	4.00	0.09	0.02	3.81	24.97
idiovol	Idiosyncratic return volatility	0.01	0.26	0.06	0.06	1.47	2.70
ill	Illiquidity	0.00	0.00	0.00	0.00	14.63	355.90
indmom	Industry momentum	-1.00	3.56	0.14	0.12	1.26	5.39
invest	Capital expenditures and inventory	-0.52	2.21	0.08	0.04	2.51	12.80
lev	Leverage	0.00	77.75	2.28	0.69	5.47	43.92
Meanrec	Mean number of analysts	1.00	4.50	2.22	2.20	-0.01	-0.32
mom12m	12-month momentum	-1.00	11.60	0.13	0.06	2.89	21.73
mom1m	1-month momentum	-0.70	2.11	0.01	0.00	1.16	7.77
mom36m	36-month momentum	-0.98	16.20	0.33	0.16	3.08	20.08
ms	Financial statement score	0.00	8.00	3.73	4.00	-0.03	-0.72
mve	Size	6.02	18.90	11.77	11.63	0.29	-0.31
mve_ia	Industry-adjusted size	-16,608.51	133,635.00	-158.25	-359.12	9.17	120.31
nanalyst	Number of analysts covering stock	0.00	34.00	5.17	3.00	1.75	2.80
nincr	Number of earnings increases	0.00	8.00	1.00	1.00	2.15	6.35
obklg		0.00	4.59	0.40	0.19	2.99	10.85
orgcap	Organizational capital	0.00	0.18	0.01	0.01	2.73	11.60
pchcapx_ia	Industry adjusted % change in capital ependitures	-237.42	1640.09	6.50	-0.35	15.17	273.34
pchcurrat	% change in current ration	-0.89	6.72	0.06	-0.01	3.82	23.69
pchdepr	% change in depreciation	-0.85	7.37	0.10	0.03	4.56	36.69
pchgm_pchsale	% change in gross margin - % change in sales	-12.26	4.77	-0.06	0.00	-5.49	54.90
pchsale_pchinvt	% change in sales - % change in inventory	-11.61	3.02	-0.06	0.01	-5.85	57.26
pchsale_pchrect	% changes in sales - % change in A/R	-7.93	3.11	-0.04	0.00	-2.90	24.02
pchsale_pchxsga	% change in sales - % change in SG&A	-3.50	4.34	0.02	0.00	3.42	31.72
pctacc	Percent accruals	-64.75	71.43	-0.65	-0.27	-1.90	34.80
pricedelay	Price delay	-15.85	15.52	0.14	0.06	0.09	39.90
ps	Financial statement score	0.00	8.00	4.18	4.00	0.03	-0.55
rd_mve	R&D to market capitalization	0.00	2.23	0.06	0.03	5.12	48.03
rd_sale	R&D to sales	0.00	283.48	0.61	0.03	23.58	733.49
retvol	Return volatility	0.00	0.27	0.03	0.02	2.42	8.82
roaq	Return on assets	-0.48	0.16	0.00	0.01	-3.40	16.22

roavol	Earnings volatility	0.00	0.85	0.03	0.01	5.31	41.10
roe	Return on equity	-7.05	8.80	0.03	0.10	-2.56	35.80
roeq	Quarterly return on equity	-2.22	1.66	0.00	0.02	-2.80	29.85
roic	Return on invested capital	-21.24	1.01	-0.08	0.07	-10.40	149.98
rsup	Revenue surprise	-4.51	2.33	0.02	0.01	-3.83	64.41
salecash	Sales to cash	0.00	2,503.48	52.60	10.60	7.68	73.99
saleinv	Sales to inventory	0.29	1,031.22	25.91	7.59	6.96	62.95
salerec	Sales to receivables	0.00	594.00	11.68	5.94	5.25	31.16
sfe	Sscaled earnings forecast	-36.23	1.09	-0.06	0.04	-14.77	296.41
sgr	Sales growth	-0.91	8.50	0.18	0.09	5.68	49.16
sp	Sales to price	0.00	54.59	2.32	1.10	4.51	29.91
spi	Industry-adjusted sales to price	-0.66	0.19	-0.01	0.00	-5.20	40.89
std_dolvol	Volatility of liquidity (dollar trading volume)	0.18	2.74	0.86	0.79	0.75	0.18
std_turn	Volatility of liquidity (share turnover)	0.02	184.01	3.90	1.90	6.07	64.87
stdcf	Cash flow volatility	0.00	1,882.88	9.88	0.14	11.94	178.00
sue	Unexpected quarterly earnings	-5.20	1.70	0.00	0.00	-13.43	550.18
tang	Debt capacity/firm tangibility	0.04	0.98	0.54	0.55	-0.14	0.99
tb	Tax income to book income	-27.70	15.36	-0.10	-0.03	-4.47	66.74
turn	Share turnover	0.00	195.94	1.02	0.52	20.43	1,384.13
zerotrade	Zero trading days	0.00	19.95	1.31	0.00	3.03	9.22

Table 2: Descriptive statistics of candidate factors

This table presents descriptive statistics of the candidate factors we apply in this study. Our sample period is from 1962 to 2018. The candidate factors are constructed in a similar vein as Hou et al. (2018). We use the same screening criteria, delisting procedure, and period similar to what they do. The first column presents the identification numbers and names of the candidate factors according to their papers. The second column presents the paper introducing these candidate factors. The last four columns present the number of observations, the mean of candidate factors, t-stat testing the mean is statistically different from zero, and the standard deviation of candidate factors. All candidate factors are based on 1-month calculation, and these portfolios are equal-weighted returns. ***, **, and * present 1%, 5%, and 10% significance level.

Candidate factors	Reference papers	# obs	mean	t-stat	std.dev
	A. Momentum				
A.1.1 Standardized unexpected earnings	Foster, Olsen, and Shevlin (1984)	534	0.009	4.966***	0.04
A.1.2 Cumulative abnormal returns around earnings announcement dates	Chan, Jegadeesh, and Lakonishok (1996)	521	0.016	8.571***	0.043
A.1.4 Price momentum, prior 6-month returns	Jegadeesh and Titman (1993)	534	0.01	3.126***	0.077
A.1.5 Price momentum, prior 11-month returns	Fama and French (1996)	534	0.014	4.244***	0.077
A.1.6 Industry momentum	Moskowitz and Grinblatt (1999)	534	0.569	2.227**	5.905
A.1.7 Revenue surprises	Jegadeesh and Livnat (2006)	534	0.002	1.371	0.037
A.1.10 The number of quarters with consecutive earnings increase	Barth, Elliott, and Finn (1999)	533	0.005	1.758*	0.071
A.1.11 52-week high	George and Hwang (2004)	529	-0.001	-0.181	0.068
A.1.12 Residual momentum, prior 6-month returns	Blitz, Huij, and Martens (2011)	534	0.003	1.398	0.056
A.1.13 Residual momentum, prior 11-month returns	Blitz, Huij, and Martens (2011)	534	0.01	3.887***	0.06
	B. Value versus growth				
B.2.1 Book-to-market equity	Rosenberg, Reid, and Lanstein (1985)	534	0.005	2.170**	0.05
B.2.2 Book-to-June-end market equity	Asness and Frazzini (2013)	534	0.005	2.327**	0.053
B.2.3 Quarterly book-to-market equity		534	0.018	6.875***	0.061
B.2.6 Assets-to-market	Rosenberg, Reid, and Lanstein (1985)	534	0.005	2.066**	0.056
B.2.8 Reversal.	De Bondt and Thaler (1985)	534	-0.004	-1.808*	0.056
B.2.9 Earnings-to-price	Basu (1983)	534	0.002	0.933	0.057
B.2.12 Cash flow-to-price	Lakonishok, Shleifer, and Vishny (1994)	534	0	0.123	0.048
B.2.14 Dividend yield	Litzenberger and Ramaswamy (1979)	534	0.002	1.008	0.039
B.2.16 Payout yield	Boudoukh et al. (2007)	529	0.005	2.587**	0.046
B.2.16 Net payout yield	Boudoukh et al. (2007)	529	0.005	2.506**	0.048
B.2.18 5-year sales growth rank	Lakonishok, Shleifer, and Vishny (1994)	534	-0.001	-0.715	0.044
B.2.19 Sales growth	Lakonishok, Shleifer, and Vishny (1994)	534	-0.002	-1.132	0.042
B.2.20 Enterprise multiple	Loughran and Wellman (2011)	534	-0.005	-2.000**	0.056

B.2.22 Sales-to-price	Barbee, Mukherji, and Raines (1996)	534	0.007	2.683**	0.058
B.2.26 Intangible return	Daniel and Titman (2006)	534	-0.009	-4.884***	0.044
B.2.30 Equity duration	Dechow, Sloan, and Soliman (2004)	534	-0.008	-3.180***	0.056
		C. Investment			
C.3.1 Abnormal corporate investment	Titman, Wei, and Xie (2004)	534	-0.003	-2.114**	0.031
C.3.2 Investment-to-assets	Cooper, Gulen, and Schill (2008)	534	0.002	4.031***	0.011
C.3.3 Quarterly investment-to-assets		522	-0.001	-0.672	0.026
C.3.4 Changes in PPE and inventory-to-assets	Lyandres, Sun, and Zhang (2008)	534	-0.004	-3.012***	0.033
C.3.5 Noa and dNoa, (changes in) net operating assets	Hirshleifer et al. (2004)	534	-0.006	-4.058***	0.032
C.3.6 Changes in long-term net operating assets.	Fairfield, Whisenant, and Yohn (2003)	534	-0.004	-3.005**	0.034
C.3.7 Investment growth	Xie (2008)	534	-0.004	-3.485***	0.028
C.3.8 2-year investment growth	Anderson and Garcia-Feijoo (2006)	534	-0.003	-1.930**	0.032
C.3.9 3-year investment growth	Anderson and Garcia-Feijoo (2006)	534	-0.002	-1.383	0.034
C.3.10 Net stock issues	Pontiff and Woodgate (2008)	534	-0.004	-3.554***	0.026
C.3.11 Percentage change in investment relative to industry	Abarbanell and Bushee (1998)	534	-0.003	-2.336**	0.034
C.3.12 Composite equity issuance	Daniel and Titman (2006)	534	-0.002	-0.817	0.044
C.3.13 Composite debt issuance	Lyandres, Sun, and Zhang (2008)	534	-0.001	-0.425	0.038
C.3.14 Inventory growth	Belo and Lin (2011)	534	-0.003	-2.061**	0.033
C.3.15 Inventory changes	Thomas and Zhang (2002)	534	-0.004	-2.922***	0.031
C.3.16 Operating accruals	Sloan (1996)	534	-0.003	-2.169**	0.033
C.3.17 Total accruals	Richardson, Sloan, Soliman, and Tuna (2005)	534	-0.003	-1.962*	0.035
C.3.18 Changes in net noncash working capital, in current operating assets, and in current operating liabilities.	Richardson, Sloan, Soliman, and Tuna (2005)	534	-0.002	-1.118	0.036
C.3.19 Changes in noncurrent operating assets	Richardson, Sloan, Soliman, and Tuna (2005)	534	-0.005	-3.415***	0.032
C.3.19 Changes in noncurrent operating liabilities	Richardson, Sloan, Soliman, and Tuna (2005)	534	-0.001	-0.867	0.03
C.3.19 Changes in net noncurrent operating assets	Richardson, Sloan, Soliman, and Tuna (2005)	534	-0.004	-3.340***	0.03
C.3.20 Changes in book equity	Richardson, Sloan, Soliman, and Tuna (2005)	534	-0.001	-0.285	0.051
C.3.20 Changes in net financial assets	Richardson, Sloan, Soliman, and Tuna (2005)	534	0.002	2.040**	0.028
C.3.20 Changes in financial liabilities	Richardson, Sloan, Soliman, and Tuna (2005)	534	-0.001	-1.339	0.023
C.3.20 Changes in in long-term investments	Richardson, Sloan, Soliman, and Tuna (2005)	534	-0.001	-1.357	0.025
C.3.20 Changes in short-term investments	Richardson, Sloan, Soliman, and Tuna (2005)	534	0	0.393	0.024
C.3.21 Discretionary accruals computed from Nasdaq Index	Xie (2001)	516	-0.003	-1.936*	0.036
C.3.21 Discretionary accruals computed from NYSE and Amex	Xie (2001)	534	-0.002	-1.407	0.03
C.3.22 Percent operating accruals	Hafzalla, Lundholm, and Van Winkle (2011)	534	-0.004	-3.059***	0.031

C.3.23 Percent total accruals	Hafzalla, Lundholm, and Van Winkle (2011)	534	-0.002	-1.417	0.026
C.3.24 Percent discretionary accruals	Hafzalla, Lundholm, and Van Winkle (2011)	534	-0.003	-2.314**	0.034
C.3.25 Net debt financing	Bradshaw, Richardson, and Sloan (2006), 1972/7	528	-0.002	-1.942*	0.026
C.3.25 Net equity financing	Bradshaw, Richardson, and Sloan (2006), 1972/8	528	-0.002	-0.797	0.047
C.3.25 Net external financing	Bradshaw, Richardson, and Sloan (2006), 1972/9	528	-0.003	-1.834*	0.042
D. Profitability					
D.4.1 Return on equity	Hou, Xue, and Zhang (2015)	534	0.018	8.135***	0.051
D.4.2 4-quarter change in return on equity		528	0.004	2.715**	0.035
D.4.3 Roa1, Roa6, and Return on assets	Balakrishnan, Bartov, and Faurel (2010), 1972/1	534	0.015	7.404***	0.048
D.4.4 4-quarter change in return on assets.		522	0.005	2.814***	0.037
D.4.5 Assets turnover	Soliman (2008)	534	0.001	0.54	0.045
D.4.5 Profit margin	Soliman (2008)	534	0.001	0.244	0.049
D.4.5 Return on net operating assets	Soliman (2008)	534	0.001	0.485	0.043
D.4.6 Capital turnover	Haugen and Baker (1996)	534	0.001	0.891	0.039
D.4.7 Quarterly assets turnover		534	0.004	2.176**	0.043
D.4.7 Quarterly profit margin		534	0.005	2.429**	0.048
D.4.7 Quarterly return on net operating assets		486	0.005	2.374**	0.045
D.4.8 Quarterly capital turnover	Haugen and Baker (1996)	534	0.006	3.634***	0.041
D.4.9 Gross profits-to-assets.	Novy-Marx (2013)	534	0.003	1.902*	0.034
D.4.10 Gross profits-to-lagged assets		534	0	0.196	0.036
D.4.11 Quarterly gross profits-to-lagged assets	Fama and French (2015)	486	0.004	3.141***	0.031
D.4.12 Operating profits to equity	Fama and French (2015)	534	0.002	1.14	0.05
D.4.13 Operating profits-to-lagged equity		534	0.001	0.4	0.045
D.4.14 Quarterly operating profits-to-lagged equity		534	0.008	3.392***	0.055
D.4.15 Operating profits-to-assets	Ball et al. (2015)	534	0.004	2.037**	0.044
D.4.16 Operating profits-to-lagged assets		534	0.003	1.411	0.043
D.4.17 Quarterly operating profits-to-lagged assets		486	0.009	4.303***	0.044
D.4.18 Cash-based operating profitability	Ball et al. (2016)	534	0.007	3.530***	0.043
D.4.19 Cash-based operating profits-to-lagged assets		534	0.005	2.761***	0.041
D.4.20 Quarterly cash-based operating profits-to-lagged assets		486	0.007	4.324***	0.035
D.4.21 Fundamental score.	Piotroski (2000)	528	0.002	1.700*	0.033
D.4.24 Ohlson's O-score	Dichev (1998)	534	0.001	0.317	0.043
D.4.25 Quarterly O-score		486	-0.002	-1.26	0.034
D.4.26 Altman's Z-score	Dichev (1998)	534	-0.004	-2.011**	0.046
D.4.27 Quarterly Z-score		486	-0.005	-2.072**	0.052
D.4.29 Taxable income-to-book income.	Lev and Nissim (2004)	534	0	0.244	0.03
D.4.30 Quarterly taxable income-to-book income		534	0.001	0.581	0.038
D.4.31 Growth score	Mohanram (2005)	348	0.004	1.079	0.077
D.4.32 Book leverage	Fama and French (1992)	534	0.001	0.438	0.039
D.4.33 Quarterly book leverage		534	0	0.136	0.042
E. Intangibles					
E.5.1 Industry adjusted organizational capital-to-assets	Eisfeldt and Papanikolaou (2013)	534	0.001	0.353	0.042

E.5.2 Advertising expense-to-market	Chan, Lakonishok, and Sougiannis (2001)	534	0	0.171	0.029
E.5.3 Growth in advertising expense.	Lou (2014)	534	0.002	3.384***	0.014
E.5.4 R&D expense-to-market	Chan, Lakonishok, and Sougiannis (2001)	534	-0.002	-0.933	0.045
E.5.8 Operating leverage	Novy-Marx (2011)	534	0.001	0.438	0.033
E.5.9 Olq1, Olq6, and Olq12, quarterly operating leverage	Novy-Marx (2011)	522	0.004	2.549**	0.032
E.5.10 Hiring rate	Belo, Lin, and Bazdresch (2014)	534	0.002	2.939***	0.014
E.5.11 R&D capital-to-assets	Li (2011)	534	0	0.254	0.041
E.5.12 Bca, brand capital-to-assets.	Belo, Lin, and Vitorino (2014)	516	0.007	2.046**	0.074
E.5.17 Ha, industry concentration (assets)	Hou and Robinson (2006)	534	-0.003	-1.248	0.047
E.5.17 He, industry concentration (book equity)	Hou and Robinson (2006)	534	-0.002	-1.098	0.044
E.5.17 Hs, industry concentration (sales)	Hou and Robinson (2006)	534	-0.003	-1.394	0.043
E.5.19 D1, price delay	Hou and Moskowitz (2005)	534	0.002	0.976	0.043
E.5.19 D2, price delay	Hou and Moskowitz (2005)	534	0	-0.107	0.023
E.5.19 D3, price delay	Hou and Moskowitz (2005)	534	0	-0.405	0.023
E.5.20 % change in sales minus % change in inventory	Abarbanell and Bushee (1998)	534	0	0.438	0.004
E.5.21 % change in sales minus % change in accounts receivable	Abarbanell and Bushee (1998)	534	0	1.085	0.006
E.5.22 % change in gross margin minus % change in sales	Abarbanell and Bushee (1998)	534	0.001	2.281**	0.006
E.5.23 % change in sales minus % change in SG&A	Abarbanell and Bushee (1998)	534	0	1.425	0.005
E.5.24 Effective tax rate	Abarbanell and Bushee (1998)	534	0	1.425	0.005
E.5.25 Labor force efficiency	Abarbanell and Bushee (1998)	534	0	1.173	0.005
E.5.26 Analysts coverage	Elgers, Lo, and Pfeiffer (2001)	485	-0.001	-0.432	0.03
E.5.27 Tangibility	Hahn and Lee (2009)	534	-0.001	-0.828	0.026
E.5.28 Quarterly tangibility.	Hahn and Lee (2009)	534	0	0.164	0.033
E.5.29 Industry-adjusted real estate ratio	Tuzel (2010)	534	0.001	0.467	0.037
E.5.30 Financial constraints (the Kaplan-Zingales index)	Lamont, Polk, and Saa-Requejo (2001)	534	0.002	1.521	0.03
E.5.32 Financial constraints (the Whited-Wu index)	Whited and Wu (2006)	534	0	0.125	0.028
E.5.33 Wwq1, Wwq6, and Wwq12, the quarterly Whited-Wu index	Whited and Wu (2006)	534	0.001	0.33	0.039
E.5.34 Secured debt-to-total debt	Valta (2016)	534	-0.001	-0.646	0.03
E.5.35 Convertible debt-to-total debt	Valta (2016)	534	0.001	0.826	0.042
E.5.37 Cta1, Cta6, and Cta12, cash-to-assets	Palazzo (2012)	534	0.002	1.079	0.045
E.5.41 Earnings persistence	Rajgopal, Shevlin, and Venkatachalam (2003)	534	-0.001	-0.665	0.032
E.5.41 Earnings predictability	Rajgopal, Shevlin, and Venkatachalam (2003)	534	-0.004	-2.162**	0.041
E.5.42 Earnings smoothness	Francis et al. (2004)	534	-0.001	-1.012	0.027
E.5.44 Earnings conservatism	Francis et al. (2004)	534	-0.002	-1.479	0.027
E.5.44 Earnings timeliness	Francis et al. (2004)	534	0	0.103	0.032
E.5.44 Earnings conservatism	Francis et al. (2004)	534	0.001	0.757	0.019
E.5.44 Earnings timeliness	Francis et al. (2004)	534	0.001	1.108	0.022
E.5.45 FRM, Pension plan funding rate	Franzoni and Martin (2006)	534	0.001	0.977	0.024

E.5.45 FRA, Pension plan funding rate	Franzoni and Martin (2006)	534	-0.002	-1.695*	0.03
E.5.46 Ala, asset liquidity	Ortiz-Molina and Phillips (2014)	486	0.000	-0.122	0.045
E.5.46 Alm, asset liquidity	Ortiz-Molina and Phillips (2014)	486	0.004	1.692	0.051
E.5.51 Average returns Ra1	Heston and Sadka (2008)	534	0.001	7.550***	0.002
E. 5.51 Average returns Ra[2,5]	Heston and Sadka (2008)	534	0.001	3.604***	0.004
E.5.51 Average returns Ra[6,10]	Heston and Sadka (2008)	534	0.000	3.554***	0.003
E.5.51 Average returns Rn1	Heston and Sadka (2008)	534	0.002	4.953***	0.007
E. 5.51 Average returns Rn[2,5]	Heston and Sadka (2008)	534	0.002	3.354***	0.012
E.5.51 Average returns Rn[6,10]	Heston and Sadka (2008)	534	0.001	3.143***	0.009
E.5.51 Average returns Rn[16,20]	Heston and Sadka (2008)	534	0.002	1.152	0.044
F. Trading frictions					
F.6.1 Me, market equity	Banz (1981)	534	-0.001	-0.485	0.05
F.6.2 Ivff1, Ivff6, and Ivff12, idiosyncratic volatility per the Fama and French (1993) 3-factor model	Ang, Hodrick, Xing, and Zhang (2006)	534	-0.01	-2.542**	0.088
F.6.3 Iv, idiosyncratic volatility	Ali, Hwang, and Trombley (2003)	534	-0.012	-3.406***	0.079
F.6.5 Ivq1, Ivq6, and Ivq12, idiosyncratic volatility		534	-0.011	-3.340***	0.079
F.6.6 Tv1, Tv6, and Tv12, total volatility	Ang, Hodrick, Xing, and Zhang (2006)	534	-0.013	-3.427***	0.089
F.6.8 β1, β6, and β12, market beta	Fama and MacBeth (1973)	534	0.000	-0.125	0.08
F.6.9 βFP1, βFP6, and βFP12, the Frazzini-Pedersen beta	Frazzini and Pedersen (2013)	534	-0.006	-1.529	0.095
F.6.10 βD1, βD6, and βD12, the Dimson beta	Dimson (1979)	533	-0.001	-0.51	0.058
F.6.11 Tur1, Tur6, and Tur12, share turnover	Datar, Naik, and Radcliffe (1998)	534	-0.002	-0.823	0.063
F.6.12 Cvt1, Cvt6, and Cvt12, coefficient of variation of share turnover	Chordia, Subrahmanyam, and Anshuman (2001)	533	0.000	-0.106	0.034
F.6.13 Dtv1, Dtv6, and Dtv12, dollar trading volume	Brennan, Chordia, and Subrahmanyam (1998)	533	-0.001	-0.605	0.034
F.6.14 Cvd1, Cvd6, and Cvd12, coefficient of variation of dollar trading volume.	Chordia, Subrahmanyam, and Anshuman (2001)	533	0.001	0.37	0.033
F.6.15 Pps1, Pps6, and Pps12, share price	Miller and Scholes (1982)	534	0.000	0.127	0.084
F.6.16 Ami1, Ami6, and Ami12, absolute return-to-volume	Amihud (2002)	533	-0.001	-0.396	0.05
F.6.17 Lm11, Lm16, Lm112, turnover-adjusted number of zero daily volume	Liu (2006)	533	0.000	-0.014	0.058
F.6.17. Lm121, Lm126, Lm1212, turnover-adjusted number of zero daily volume	Liu (2006)	533	0.002	0.695	0.06
F.6.17, Lm61, Lm66, Lm612, turnover-adjusted number of zero daily volume	Liu (2006)	533	0.002	0.711	0.061
F.6.18 Mdr1, Mdr6, and Mdr12, maximum daily return	Bali, Cakici, and Whitelaw (2011)	534	-0.008	-2.586**	0.074
F.6.20 Isc1, Isc6, and Isc12, idiosyncratic skewness per the CAPM		534	0.003	2.249**	0.027
F.6.21 Isff1, Isff6, and Isff12, idiosyncratic skewness per the Fama and French	Bali, Engel, and Murray (2016)	534	0.003	2.756***	0.025

F.6.23 Cs1, Cs6, and Cs12, coskewness	Harvey and Siddique (2000)	534	-0.001	-0.806	0.032
F.6.25 βlcc1, βlcc6, βlcc12, liquidity betas illiquidity-illiquidity	Kelly and Jiang (2014)	533	0.026	9.205***	0.065
F.6.25 βlcr1, βlcr6, βlcr12, liquidity betas (illiquidity-return)	Kelly and Jiang (2014)	533	0.001	0.438	0.042
F.6.25 βlrc1, βlrc6, βlrc12, liquidity betas return illiquidity	Kelly and Jiang (2014)	533	-0.003	-1.632	0.047
F.6.25 βnet1, βnet6, and βnet12, liquidity betas (net)	Kelly and Jiang (2014)	533	0.006	1.864*	0.077
F.6.25 βret1, βret6, and βret12, liquidity betas (return-return)	Kelly and Jiang (2014)	533	0.006	1.886*	0.078
F.6.26 Short-term reversal	Jegadeesh (1990)	533	0.003	1.307	0.051
F.6.27 β−1, β−6, and β−12, downside beta	Ang, Chen, and Xing (2006)	533	-0.002	-0.626	0.073
F.6.31 βPS1, βPS6, and βPS12, the Pastor-Stambaugh beta		534	0.001	0.304	0.04

Table 3 Latent factor and candidate factor selection

This table presents the results of various applications of the SOFAR method. In Panel A, we run the regression (equation (2)) without imposing any tuning parameters. The only restriction is to select 5 latent factors with the highest eigenvalues. The panel presents the eigenvalues of these latent factors. Each latent factor is a linear combination of candidate factors presented in Table 2. We also present the coefficients of each candidate factors in each latent factor. Panel B tunes in λ_a and λ_d and Panel C tunes in λ_a, λ_b, and λ_d. Eigenvalues associated with factors and the coefficients of the candidate factors associated with each latent factor are presented. Each latent factor is correlated with only a fraction of testing assets. Panel C also presents the number of assets associated with each latent factor. The sample period is 1972 to 2018. The testing assets are 202 portfolios described in Section 4. We standardize the candidate factors and testing assets.

Panel A: No tuning parameter adjustment

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Eigenvalues	17.82	3.74	2.01	1.55	1.29
Coefficients of candidate factors associated with the latent factors					
A.1 Momentum					
A.1.1 Standardized unexpected earnings	-0.03	0.02	-0.05	-0.03	-0.03
A.1.2 Cumulative abnormal returns around earnings announcement date	0.01	0.02	-0.03	0.01	-0.01
A.1.4 Prior 6-month returns	0.05	0.02	0.06	0.09	0.07
A.1.5 Prior 11-month returns	0.00	0.04	0.00	0.18	0.03
A.1.6 Industry momentum	-0.02	0.01	0.01	0.00	-0.02
A.1.7 Revenue surprises	0.04	0.00	0.02	-0.01	-0.02
A.1.10 The number of quarters with consecutive earnings increase	0.03	-0.03	-0.01	-0.02	0.00
A.1.11 52-week high	0.11	-0.04	0.05	-0.01	-0.03
A.1.12 6-month residual momentum	-0.08	-0.04	-0.04	-0.04	0.04
A.1.13 11-month residual momentum	0.03	-0.01	0.09	0.24	0.19
A.2 Value versus growth					
A.2.1 Book-to-market equity	0.05	-0.03	0.01	-0.02	0.02
A.2.2 Book-to-June-end market equity	0.00	0.00	0.07	0.02	-0.08
A.2.3 Quarterly book-to-market equity	0.04	-0.09	-0.11	-0.15	-0.10
A.2.6 Assets-to-market	-0.10	-0.01	0.03	0.09	-0.01
A.2.8 Reversal	-0.01	0.05	-0.01	0.03	0.00
A.2.9 Earnings-to-price	0.09	0.01	-0.02	-0.02	0.03
A.2.12 Cash flow-to-price	-0.08	0.01	-0.02	0.00	-0.01
A.2.14 Dividend yield	0.18	-0.07	-0.11	-0.02	-0.08
A.2.16 Net payout yield	0.00	-0.01	-0.09	-0.05	0.12
A.2.16 Payout yield	-0.02	0.02	-0.03	0.02	0.05
A.2.18 Sales growth rank	0.00	-0.06	0.09	-0.04	0.10
A.2.19 Sales growth	0.00	0.00	-0.01	-0.01	-0.06

A.2.20 Enterprise multiple	-0.01	0.03	0.11	0.05	-0.01
A.2.22 Sales-to-price	-0.03	-0.01	-0.01	-0.12	-0.08
A.2.26 Intangible return	0.08	-0.01	-0.05	0.05	-0.02
A.2.30 Equity duration	0.03	0.03	0.02	-0.07	-0.02
A.3 Investment					
C.3.1 Abnormal corporate investment	-0.01	-0.02	-0.01	0.01	-0.01
C.3.2 Investment-to-assets	0.11	0.05	0.04	-0.21	-0.12
C.3.3 Quarterly investment-to-assets	0.00	-0.02	0.02	0.01	0.06
C.3.4 Changes in PPE and inventory-to-assets	-0.02	0.02	0.04	0.02	-0.07
C.3.5 Changes in net operating assets	-0.01	-0.03	0.05	-0.03	-0.03
C.3.6 Changes in long-term net operating assets.	0.11	0.02	0.02	-0.04	0.03
C.3.7 Investment growth	0.05	-0.01	0.00	-0.01	-0.06
C.3.8 2-year investment growth	-0.03	0.03	-0.01	-0.05	0.03
C.3.9 3-year investment growth	0.04	0.00	0.01	0.01	0.00
C.3.10 Net stock issues	0.00	-0.01	-0.01	-0.04	-0.02
C.3.11 Percentage change in investment relative to industry	0.00	0.01	0.01	0.05	0.08
C.3.12 Composite equity issuance	0.00	0.36	0.01	-0.12	0.10
C.3.13 Composite debt issuance	0.02	0.01	-0.02	-0.05	-0.02
C.3.14 Inventory growth	0.01	0.02	-0.01	-0.04	-0.04
C.3.15 Inventory changes	-0.02	-0.01	-0.03	0.05	0.01
C.3.16 Operating accruals	0.03	0.08	-0.02	-0.02	-0.06
C.3.17 Total accruals	0.02	0.00	-0.01	0.03	0.04
C.3.18 Changes in net noncash working capital, in current operating assets, and in current operating liabilities.	0.01	0.04	0.02	0.01	0.01
C.3.19 Changes in noncurrent operating assets	0.01	0.01	0.03	-0.01	-0.01
C.3.19 Changes in noncurrent operating liabilities	0.01	-0.01	0.00	-0.02	0.03
C.3.19 Changes in net noncurrent operating assets	-0.08	0.01	-0.07	0.02	0.02
C.3.20 Changes in book equity	0.26	0.57	-0.20	0.01	-0.08
C.3.20 Changes in net financial assets	0.01	-0.01	-0.02	-0.04	0.05
C.3.20 Changes in financial liabilities	0.00	-0.01	0.01	0.00	0.03
C.3.20 Changes in in long-term investments	0.02	-0.01	-0.02	0.02	-0.05
C.3.20 Changes in short-term investments	0.04	0.02	0.03	0.01	-0.03
C.3.21 Discretionary accruals computed from Nasdaq Index	0.01	0.00	0.02	-0.01	0.00
C.3.21 Discretionary accruals computed from NYSE and Amex	0.00	-0.03	0.00	0.02	0.01
C.3.22 Percent operating accruals	-0.01	-0.04	0.00	-0.06	-0.04
C.3.23 Percent total accruals	-0.02	0.02	0.00	0.03	-0.05
C.3.24 Percent discretionary accruals	-0.02	-0.01	-0.02	0.03	0.03
C.3.25 Net debt financing	-0.03	0.01	-0.04	0.06	-0.06
C.3.25 Net equity financing	-0.04	-0.02	-0.04	0.07	-0.17
C.3.25 Net external financing	-0.06	0.02	0.01	-0.04	0.02
A.4 Profitability					

D.4.1 Return on equity	0.11	0.04	0.01	0.00	0.07
D.4.2 4-quarter change in return on equity	-0.03	-0.02	0.04	-0.03	0.01
D.4.3 Return on assets	-0.10	-0.07	0.00	0.02	-0.04
D.4.4 4-quarter change in return on assets.	0.02	0.03	0.01	0.01	0.00
D.4.5 Assets turnover	0.07	0.00	0.17	-0.04	0.09
D.4.5 Profit margin	0.00	0.06	-0.04	-0.09	0.05
D.4.5 Return on net operating assets	0.00	-0.02	0.12	0.10	0.08
D.4.6 Capital turnover	0.06	0.09	0.05	-0.10	-0.06
D.4.7 Quarterly assets turnover	-0.04	-0.07	-0.02	0.03	0.03
D.4.7 Quarterly profit margin	0.00	0.00	-0.04	0.05	0.06
D.4.8 Quarterly capital turnover	-0.01	0.00	-0.06	0.06	0.14
D.4.9 Gross profits-to-assets	-0.04	-0.03	0.00	0.04	-0.01
D.4.10 Gross profits-to-lagged assets	-0.03	0.02	-0.05	-0.04	0.07
D.4.12 Operating profits to equity	-0.19	-0.13	0.10	-0.10	-0.07
D.4.13 Operating profits-to-lagged equity	0.01	-0.01	-0.02	-0.03	0.05
D.4.14 Quarterly operating profits-to-lagged equity	0.03	0.00	0.02	0.01	-0.02
D.4.15 Operating profits-to-assets.	-0.02	0.17	0.08	0.09	0.01
D.4.16 Operating profits-to-lagged assets	0.08	-0.08	-0.11	-0.05	0.09
D.4.18 Cash-based operating profitability	0.17	-0.03	-0.07	-0.06	0.24
D.4.19 Cash-based operating profits-to-lagged assets.	-0.14	0.09	-0.08	0.13	-0.29
D.4.21 Fundamental score.	-0.01	-0.01	0.02	0.03	0.02
D.4.24 Ohlson's O-score	0.04	0.08	-0.03	0.03	-0.02
D.4.26 Altman's Z-score	0.05	0.05	0.04	-0.01	-0.04
D.4.29 Taxable income-to-book income.	0.01	-0.02	0.06	-0.02	0.07
D.4.30 Quarterly taxable income-to-book income	0.05	-0.04	0.01	-0.08	0.06
D.4.32 Book leverage	0.09	0.02	0.08	-0.11	0.02
D.4.33 Quarterly book leverage	-0.12	0.05	-0.04	0.11	0.03
			A.5 Intangibles		
E.5.1 (Industry-adjusted) organizational capital-to-assets	0.04	-0.06	0.02	-0.06	0.08
E.5.2 Advertising expense-to-market	-0.05	0.00	0.05	-0.14	0.15
E.5.3 Growth in advertising expense.	-0.10	0.00	-0.01	0.03	0.03
E.5.4 R&D expense-to-market	0.11	0.06	0.24	-0.22	-0.06
E.5.8 Operating leverage	0.04	0.02	-0.10	0.09	0.14
E.5.9 Quarterly operating leverage	-0.02	-0.02	-0.03	-0.01	-0.08
E.5.10 Hiring rate	-0.13	-0.10	-0.05	-0.13	0.15
E.5.12 Brand capital-to-assets.	0.01	0.00	-0.03	0.03	-0.03
E.5.17 Industry concentration (assets)	-0.02	0.02	-0.11	0.01	-0.04
E.5.17 Industry concentration (book equity)	0.05	0.01	0.02	-0.02	0.05
E.5.17 Industry concentration (sales)	-0.01	-0.04	0.03	-0.02	0.08
E.5.19 D1, price delay	0.02	0.01	-0.03	0.08	0.10
E.5.19 D2, price delay	0.00	-0.04	0.03	-0.03	-0.12
E.5.19 D3, price delay	0.02	0.04	-0.04	0.02	0.05

E.5.20 % change in sales minus % change in inventory	-0.01	0.02	0.02	-0.02	0.01
E.5.21 % change in sales minus % change in accounts receivable	-0.04	0.06	-0.03	-0.06	-0.07
E.5.22 % change in gross margin minus % change in sales	0.01	-0.02	0.02	0.01	0.06
E.5.23 % change in sales minus % change in SG&A	-0.02	-0.01	0.01	0.00	-0.03
E.5.24 Effective tax rate	-0.02	-0.01	0.01	0.00	-0.03
E.5.25 Labor force efficiency	0.01	0.06	-0.02	0.09	0.01
E.5.27 Tangibility	0.15	-0.09	0.04	0.07	-0.06
E.5.28 Quarterly tangibility.	0.01	-0.02	0.02	0.00	0.03
E.5.29 Industry-adjusted real estate ratio	-0.02	-0.04	0.02	-0.13	0.01
E.5.30 Financial constraints (the Kaplan-Zingales index)	-0.13	-0.13	0.35	-0.06	-0.05
E.5.32 Financial constraints (the Whited-Wu index)	0.05	-0.05	-0.10	0.05	-0.14
E.5.33 Quarterly Whited-Wu index	-0.01	-0.01	-0.02	0.02	-0.03
E.5.34 Secured debt-to-total debt	0.08	0.00	-0.04	0.01	0.01
E.5.35 Convertible debt-to-total debt	0.05	-0.02	0.00	-0.03	-0.11
E.5.37 Cash-to-assets	-0.04	0.01	-0.01	-0.03	0.02
E.5.41 Earnings persistence	0.04	-0.01	0.01	-0.04	-0.01
E.5.41 Earnings predictability	-0.04	0.02	0.05	0.02	-0.04
E.5.42 Earnings smoothness	0.01	-0.01	0.03	-0.02	-0.02
E.5.44 Earnings conservatism.	-0.03	0.01	0.03	-0.06	-0.02
E.5.44 Earnings timeliness	-0.02	0.02	-0.01	0.05	-0.02
E.5.45 FRM, Pension plan funding rate	0.00	-0.09	0.10	0.13	0.11
E.5.45 FRA, Pension plan funding rate	0.03	0.07	-0.04	-0.07	0.05
E.5.46 Ala, asset liquidity	-0.07	-0.01	0.12	0.15	-0.15
E.5.46 Alm, asset liquidity	-0.25	-0.17	-0.01	0.00	-0.19
E.5.51 Average returns Ra1	0.00	-0.06	-0.04	-0.07	-0.01
E. 5.51 Average returns Ra25	-0.12	0.24	0.00	0.03	-0.11
E.5.51 Average returns Ra[6,11]	0.16	-0.10	0.10	0.07	0.05
E.5.51 Average returns Rn1	0.14	-0.04	0.02	0.14	0.04
E. 5.51 Average returns Rn[2,5]	0.03	-0.16	0.06	0.16	0.16
E.5.51 Average returns Rn[6,11]	-0.10	0.06	-0.11	-0.17	-0.10
E.5.51 Average returns Rn[16,20]	-0.01	0.00	-0.04	-0.05	0.03
A 6. Trading frictions					
F.6.1 Market equity	-0.14	0.03	-0.16	0.03	-0.03
F.6.2 Idiosyncratic volatility per the Fama and French (1993) 3-factor model	-0.06	0.00	0.00	-0.01	0.01
F.6.3 Idiosyncratic volatility	0.12	-0.02	0.13	-0.13	-0.14
F.6.5 Idiosyncratic volatility	-0.13	0.11	-0.02	-0.04	0.17
F.6.6 Total volatility	0.04	-0.03	0.00	0.12	-0.06
F.6.8 Market beta	-0.04	0.00	0.11	-0.18	-0.05
F.6.9 The Frazzini-Pedersen beta	-0.04	0.00	0.08	-0.09	-0.06
F.6.10 The Dimson beta	-0.03	0.02	0.02	-0.06	-0.01
F.6.11 Share turnover	0.19	0.02	0.20	0.16	-0.27
F.6.12 Coefficient of variation of share turnover	-0.10	-0.01	-0.16	-0.19	0.06

F.6.13 Dollar trading volume	-0.18	0.05	0.04	0.09	0.04
F.6.14 Coefficient of variation of dollar trading volume.	-0.01	0.10	0.14	0.19	-0.06
F.6.15 Share price	0.02	0.05	-0.08	0.09	0.07
F.6.16 Absolute return-to-volume	-0.02	-0.01	-0.08	-0.01	0.22
F.6.17 Lm1_1, Turnover-adjusted number of zero daily volume	-0.16	0.02	0.01	0.03	-0.02
F.6.17 Lm6_1, Turnover-adjusted number of zero daily volume	0.31	-0.01	0.31	-0.07	0.02
F.6.17 Lm12_1, Turnover-adjusted number of zero daily volume	0.13	-0.16	-0.17	0.13	-0.13
F.6.18 Maximum daily return	-0.08	-0.01	-0.08	-0.05	-0.04
F.6.20 Idiosyncratic skewness per the CAPM	0.00	0.01	0.00	0.03	-0.03
F.6.21 Idiosyncratic skewness per the Fama and French	0.00	-0.03	0.00	-0.04	0.02
F.6.23 Coskewness	0.01	0.00	-0.01	-0.03	-0.02
F.6.27 Downside beta	-0.07	0.04	0.09	0.03	-0.02
F.6.26 Short-term reversal	-0.02	-0.03	0.00	-0.04	0.03
F.6.25 Liquidity betas illiquidity-illiquidity	-0.09	0.07	0.11	0.06	-0.04
F.6.25 Liquidity betas (illiquidity-return)	-0.01	0.01	-0.01	-0.01	-0.03
F.6.25 Liquidity betas return illiquidity	0.04	0.00	0.02	-0.02	-0.01
F.6.25 Liquidity betas (net)	0.22	-0.22	-0.29	0.02	-0.20
F.6.25 Liquidity betas (return-return)	-0.18	0.26	0.20	0.01	0.02
F.6.31 The Pastor-Stambaugh beta	0.02	0.01	0.00	0.00	-0.05

Panel B: Tune in λ_a and λ_b

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Eigenvalues	16.46	2.80	0.91	0.81	0.77

Factor 2		Factor 3		Factor 4		Factor 5	
F.6.25 βnet1, βnet6, and βnet12, liquidity betas (net)	-0.28	E.5.30 Financial constraints (the Kaplan-Zingales index)	-0.74	F.6.11 Share turnover	-0.33	E.5.10 Hiring rate	-0.08
B.2.3 Quarterly book-to-market equity	-0.11	E.5.10 Hiring rate	-0.4	F.6.25 Liquidity betas (net)	-0.2	D.4.15 Operating profits-to-assets.	-0.07
F.6.17. Lm121, Lm126, Lm1212, turnover-adjusted number of zero daily volume	-0.04	E.5.51 Average returns Rn[2,5]	-0.23	F.6.15 Share price	-0.07	E.5.8 Operating leverage	-0.04
E. 5.51 Average returns Rn25	-0.01	E.5.4 R&D expense-to-market	-0.22	E.5.51 Average returns Ra[6,11]	-0.06	F.6.1 Market equity	-0.01
F.6.14 Cvd1, Cvd6, and Cvd12, coefficient of variation of dollar trading volume.	0.02	E.5.11 RCA, Capital-to-assets	-0.03	E.5.30 Financial constraints (the Kaplan-Zingales	-0.05	E.5.11 Capital-to-assets	0
E.5.30 Financial constraints (the Kaplan-Zingales index)	0.04	C.3.20 Changes in book equity	0.04	C.3.20 Changes in book equity	0	E.5.51 Average returns Ra[6,11]	0
F.6.25 βret1, βret6, and βret12, liquidity betas (return-return)	0.3	D.4.18 Cash-based operating profitability	0.05	F.6.25 Liquidity betas (return-return)	0.01	C.3.25 Net equity financing	0.01
F.6.25 βret1, βret6, and βret12, liquidity betas (return-return)	0.62	E.5.46 Alm, asset liquidity.	0.43	F.6.17. Lm121, Turnover-adjusted number of zero daily volume	0.07	E.5.51 Average returns Rn1	0.08
C.3.20 Changes in book equity	0.65			E.5.4 R&D expense-to-market	0.16	E.5.4 R&D expense-to-market	0.16
				F.6.17 Lm11, Turnover-adjusted number of zero daily volume	0.24	E.5.51 Average returns Rn[6,11]	0.33
				E.5.8 Operating leverage	0.53	E.5.41 Earnings persistence	0.92
				E.5.8 Operating leverage	0.68		

Panel C: Tune in λ_a. λ_b and λ_d

	Factor 1	Factor 2	Factor 3	Factor 4
Eigenvalues	15.94	2.47	0.56	0.59
Correlated Assets	202	162	135	125

Factor 2		Factor 3		Factor 4	
F.6.25 Liquidity betas (net)	-0.24	E.5.46 Alm, asset liquidity	-0.42	E.5.8 Operating leverage	-0.53
B.2.3 Quarterly book-to-market equity	-0.17	F.6.25 Liquidity betas (net)	-0.12	F.6.17 Lm11, Turnover-adjusted number of zero daily volume	-0.37
F.6.17. Lm121, Turnover-adjusted number of zero daily volume	-0.09	C.3.20 Changes in book equity	-0.09	F.6.17. Lm121, Turnover-adjusted number of zero daily volume	-0.28
F.6.25 Liquidity betas illiquidity-illiquidity	-0.01	D.4.18 Cash-based operating profitability	-0.02	D.4.12 Operating profits to equity	-0.15
E.5.46 Alm, asset liquidity.	0.00	E.5.51 Average returns Rn[2,5]	0.25	F.6.1 Market equity	-0.14
E. 5.51 Average returns Rn[2,5]	0.00	E.5.30 Financial constraints (the Kaplan-Zingales index)	0.36	D.4.32 Book leverage	-0.11
E. 5.51 Average returns Ra[2,5]	0.01	E.5.4 R&D expense-to-market	0.39	E.5.4 R&D expense-to-market	-0.06
E.5.30 Financial constraints (the Kaplan-Zingales index)	0.07	E.5.11 RCA, Capital-to-assets	0.42	C.3.20 Changes in book equity	-0.02
F.6.14 Coefficient of variation of dollar trading volume.	0.07	E.5.10 Hiring rate	0.52	F.6.8 Market beta	0.03
F.6.25 Liquidity betas (return-return)	0.27			E.5.51 Average returns Ra[6,11]	0.07
C.3.20 Changes in book equity	0.58			D.4.5 Assets turnover	0.12
C.3.12 Composite equity issuance	0.70			E.5.30 Financial constraints (the Kaplan-Zingales index)	0.18
				E.5.46 Ala asset liquidity	0.18
				F.6.25 Liquidity betas (return-return)	0.21
				F.6.3 Idiosyncratic volatility	0.27
				F.6.11 Share turnover	0.35
				F.6.25 Liquidity betas (net)	0.36

Table 4 Correlation between candidate factors

This paper presents the correlation coefficient of candidate factors associated with factors 2 in Panel A, factor 3 in Panel B, and factor 4 in Panel C. These factors are from Table 3 Panel C when λ_a. λ_b and λ_d are tuned in. For each latent factor, we present the correlation of all candidate factors with six to eight candidate factors that have the highest coefficients. The remaining correlation can be provided upon request.

Panel A: Correlation coefficients of the candidate factors that are the component of factor 1

	F.6.25 Liquidity betas (net)	B.2.3 Quarterly book-to-market equity	F.6.17. Lm121, Turnover-adjusted number of zero daily volume	F.6.25 Liquidity betas (return-return)	C.3.20 Changes in book equity	C.3.12 Composite equity issuance
F.6.25 Liquidity betas (net)	1.00	0.00	-0.73	0.98	-0.54	-0.41
B.2.3 Quarterly book-to-market equity	0.00	1.00	0.26	-0.05	0.17	0.06
F.6.17. Turnover-adjusted number of zero daily volume	-0.73	0.26	1.00	-0.72	0.60	0.50
F.6.25 Liquidity betas illiquidity-illiquidity	0.24	0.17	-0.30	0.22	-0.73	-0.71
E. 5.51 Average returns Rn[2,5]	0.57	-0.26	-0.67	0.58	-0.50	-0.38
E.5.46 Alm, asset liquidity	0.00	0.38	0.11	-0.04	-0.19	-0.28
E. 5.51 Average returns Ra[2,5]	0.52	-0.28	-0.61	0.53	-0.41	-0.30
F.6.14 Coefficient of variation of dollar trading volume.	-0.03	0.04	0.11	-0.03	0.59	0.63
E.5.30 Financial constraints (the Kaplan-Zingales index)	0.12	-0.44	-0.28	0.16	0.03	0.14
F.6.25 Liquidity betas (return-return)	0.98	-0.05	-0.72	1.00	-0.52	-0.40
C.3.20 Changes in book equity	-0.54	0.17	0.60	-0.52	1.00	0.86
C.3.12 Composite equity issuance	-0.41	0.06	0.50	-0.40	0.86	1.00

Panel B: Correlation coefficients of the candidate factors that are the component of latent factor 2

	E.5.46 Alm, asset liquidity	F.6.25 Liquidity betas (net)	C.3.20 Changes in book equity	D.4.18 Cash-based operating profitability	E. 5.51 Average returns Rn[2,5]	E.5.30 Financial constraints (the Kaplan-Zingales index)	E.5.4 R&D expense-to-market	E.5.11 R&D capital-to-assets	E.5.10 Hiring rate
E.5.46 Alm, asset liquidity.	1.00	0.00	-0.19	-0.47	-0.22	-0.94	-0.04	-0.54	-0.45

F.6.25 Liquidity betas (net)	0.00	1.00	-0.54	-0.50	0.57	0.12	0.59	0.42	0.44
C.3.20 Changes in book equity	-0.19	-0.54	1.00	0.68	-0.50	0.03	-0.60	-0.37	-0.40
D.4.18 Cash-based operating profitability	-0.47	-0.50	0.68	1.00	-0.19	0.40	-0.40	0.01	-0.10
E. 5.51 Average returns Rn[2,5]	-0.22	0.57	-0.50	-0.19	1.00	0.38	0.48	0.53	0.75
E.5.30 Financial constraints (the Kaplan-Zingales index)	-0.94	0.12	0.03	0.40	0.38	1.00	0.13	0.64	0.52
E.5.4 R&D expense-to-market	-0.04	0.59	-0.60	-0.40	0.48	0.13	1.00	0.72	0.46
E.5.11 R&D capital-to-assets	-0.54	0.42	-0.37	0.01	0.53	0.64	0.72	1.00	0.65
E.5.10 Hiring rate	-0.45	0.44	-0.40	-0.10	0.75	0.52	0.46	0.65	1.00

Panel C: Correlation coefficients of the candidate factors that are the components of latent factor 3

	E.5.8 Operating leverage	F.6.17 Lm11, Turnover-adjusted number of zero daily volume	F.6.17. Lm121, Turnover-adjusted number of zero daily volume	F.6.3 Iv, idiosyncratic volatility	F.6.11 Share turnover	F.6.25 Liquidity betas (net)
E.5.8 Operating leverage	1.00	0.27	0.27	-0.16	-0.31	-0.21
F.6.17 Lm11, Turnover-adjusted number of zero daily volume	0.27	1.00	0.98	-0.67	-0.98	-0.70

F.6.17 Lm121, Turnover-adjusted number of zero daily volume	0.27	0.98	1.00	-0.68	-0.96	-0.73
D.4.12 Operating profits to equity	0.20	0.62	0.62	-0.71	-0.61	-0.59
F.6.1 Market equity	-0.10	0.15	0.13	-0.45	-0.11	-0.09
D.4.32 Book leverage	0.33	0.42	0.42	-0.19	-0.41	-0.24
E.5.4 R&D expense-to-market	0.01	-0.62	-0.61	0.56	0.59	0.59
C.3.20 Changes in book equity	-0.01	0.61	0.60	-0.77	-0.59	-0.54
F.6.8 Market beta	-0.17	-0.75	-0.77	0.71	0.76	0.91
E.5.51 Average returns Ra[6,10]	-0.30	-0.20	-0.22	0.06	0.24	0.07
D.4.5 Assets turnover	0.15	-0.57	-0.59	0.43	0.55	0.50
E.5.46 Ala asset liquidity	-0.14	-0.73	-0.75	0.59	0.74	0.64
E.5.30 Financial constraints (the Kaplan-Zingales index)	-0.24	-0.28	-0.28	-0.01	0.28	0.12
F.6.25 Liquidity betas (return-return)	-0.24	-0.70	-0.72	0.66	0.70	0.98
F.6.3 Idiosyncratic volatility	-0.16	-0.67	-0.68	1.00	0.67	0.69
F.6.11 Share turnover	-0.31	-0.98	-0.96	0.67	1.00	0.71
F.6.25 Liquidity betas (net)	-0.21	-0.70	-0.73	0.69	0.71	1.00

Table 5 Risk premium of factors

Panel A of this Table shows the risk premium of the latent factors selected by SOFAR when λ_a. λ_b and λ_d are controlled (Panel C Table 3). The 2nd and 3rd (4th and 5th) columns present the risk premium and associated t-stat estimated from Fama-McBeth regression (Giglio and Xiu (2018)'s approach). Panel B applies Giglio and Xiu's approach and presents the risk premium and associated t-stat of the candidate factors that are the components of each factor. ***, **, and * present 1%, 5% and 10% significance levels.

Panel A

	FM		Giglio and Xiu	
	Risk Premium	T-stat	Risk Premium	T-stat
Factor 1	0.03	0.80	-0.03	-1.03
Factor 2	-0.06	-1.06	-0.05	-0.93
Factor 3	-0.05	-0.51	**-0.12****	**-1.96**
Factor 4	**-0.28****	**-2.62**	**-0.22****	**-2.16**

Panel B: The risk premium of candidate factors after applying Giglio and Xu (2018)

Factor 2	Risk Premium	T-stat	Factor 3	Risk Premium	T-stat	Factor 4	Risk Premium	T-stat
B.2.1 Book-to-market equity	-0.05	-1.20	C.3.20 Changes in book equity	-0.03	-0.68	C.3.20 Changes in book equity	-0.03	-0.68
C.3.12 Composite equity issuance	-0.04	-0.86	D.4.18 Cash-based operating profitability	-0.01	-0.31	D.4.5 Assets turnover	-0.03	-0.91
C.3.20 Changes in book equity	-0.03	-0.68	E.5.4 R&D expense-to-market	-0.05	-1.42	D.4.12 Operating profits to equity	0.05	1.36
E.5.30 Financial constraints (the Kaplan-Zingales index)	-0.08**	-2.16	E.5.10 Hiring rate	-0.04	-1.08	D.4.32 Book leverage	0.11***	2.92
E.5.46 Alm, asset liquidity.	0.09**	2.41	E.5.11 R&D capital-to-assets	-0.05	-1.86	E.5.4 R&D expense-to-market	-0.05	-1.42
E. 5.51 Average returns Rn[2,5]	-0.05	-1.31	E.5.30 Financial constraints (the Kaplan-Zingales index)	-0.08**	-2.16	E.5.8 Operating leverage	0.10***	3.29
E. 5.51 Average returns Ra[2,5]	-0.06	-1.42	E.5.46 Alm, asset liquidity	0.09**	2.41	E.5.30 Financial constraints (the	-0.08**	-2.16

						Kaplan-Zingales index)		
F.6.14 Coefficient of variation of dollar trading volume.	-0.04	-1.09	E. 5.51 Average returns Rn[2,5]	-0.05	-1.31	E.5.46 Ala asset liquidity	-0.06	-1.43
F.6.17 Lm11, Turnover-adjusted number of zero daily volume	0.06	1.39	F.6.25 Liquidity betas (net)	-0.09**	-2.20	E.5.51 Average returns Ra[6,10]	-0.07**	-2.15
F.6.25 Liquidity betas illiquidity-illiquidity	0.03	0.82				F.6.1 Market equity	-0.01	-0.35
F.6.25 Liquidity betas (net)	-0.09**	-2.20				F.6.3 Idiosyncratic volatility	-0.06	-1.59
F.6.25 Liquidity betas (return-return)	-0.09**	-2.19				F.6.8 Market beta	-0.06	-1.51
						F.6.11 Share turnover	-0.06	-1.41
						F.6.17 Lm11, turnover-adjusted number of zero daily volume	0.05	1.29
						F.6.17. Lm121, turnover-adjusted number of zero daily volume	0.06	1.39
						F.6.25 Liquidity betas (return-return)	-0.09**	-2.19
						F.6.25 Liquidity betas (net)	-0.09**	-2.20

Table 6 Robustness to Tuning parameter and different testing portfolios

In this Table, we run the regression (equation (2)), and tune in λ_a, λ_b, and λ_d. The panel presents the eigenvalues of these latent factors. Each latent factor is a linear combination of candidate factors presented in Table 2. We also present the coefficients of each candidate factors in each latent factor. Each latent factor is correlated with only a fraction of testing assets. The number of assets associated with each latent factor is also presented. There are three parameters to control the tuning parameters λ_d, λ_a and λ_a: the maximum total number of latent factors, and the upper and lower bound for the three tuning parameters. The maximum total number of latent factors provides an upper bound for the total number of latent factors. For example, we can set it equal to 10, which implies that we cannot select more than 10 latent factors, but we might only select 5 latent factors by imposing λ_d,. The maximum total number of latent factors guarantee that for any λ_d, the number of PC is less than or equal to 10. The upper and lower bound for the three tuning parameters are used to controlling the tuning parameter directly. The SOFAR automatically estimates the maximum values of the three tuning parameters. The upper and lower bounds control the percentage of the maximum parameters. For example, if the upper bound is 0.95 and the lower bound is 0.05, the tuning parameter should be between 0.05 and 0.95 of the maximum values of these parameters. Panel A sets up the maximum number of factors to 8, and Panel B change the max and min of each lambda from 0.95 and 0.05 to 0.97 and 0.03. Eigenvalues associated with five factors, the coefficients of the candidate factors associated with each factor are presented, and the number of assets correlated with each factor is presented. Coefficients of candidate factors associated with the latent factors are also shown. The sample period is 1972 to 2018. The testing assets are 202 portfolios described in Section 4. We standardize the candidate factors and testing assets.

Panel A: Maximum number of factors is eight

	Factor 1	Factor 2	Factor 3	Factor 4
Eigenvalues	16.41	2.50	0.67	0.67
Correlated Assets	202	164	144	131

Coefficients of candidate factors associated with the latent factors

Factor 2		Factor 3		Factor 4	
F.6.25 Liquidity betas (net)	-0.23	E.5.46 Asset liquidity	-0.46	E.5.8 Operating leverage	-0.53
B.2.3 Quarterly book-to-market equity	-0.19	C.3.20 Changes in book equity	-0.07	F.6.1 Market equity	-0.30
F.6.17. Turnover-adjusted number of zero daily volume	-0.11	D.4.18 Cash-based operating profitability	-0.03	F.6.17 Lm11, turnover-adjusted number of zero daily volume	-0.28
E.5.51 Average returns Rn25	-0.05	F.6.25 Liquidity betas (net)	-0.02	F.6.17. Lm121, turnover-adjusted number of zero daily volume	-0.24
F.6.25 Liquidity betas illiquidity-illiquidity	-0.04	E.5.51 Average returns Rn25	0.24	D.4.32 Book leverage	-0.10
E.5.46 Asset liquidity.	-0.01	E.5.4 R&D expense-to-market	0.25	F.6.15 Share price	-0.10
D.4.6 Capital turnover	0.00	E.5.11 Capital-to-assets	0.33	D.4.12 Operating profits to equity	-0.03
E.5.51 Average returns Ra25	0.05	E.5.10 Hiring rate	0.36	D.4.5 Assets turnover	0.00
E.5.30 Financial constraints (the Kaplan-Zingales index)	0.07	E.5.30 Financial constraints (the Kaplan-Zingales index)	0.65	E.5.30 Financial constraints (the Kaplan-Zingales index)	0.01

F.6.14 Coefficient of variation of dollar trading volume.	0.08	F.6.8 Market beta	0.17
F.6.25 Liquidity betas (return-return)	0.29	F.6.25 Liquidity betas (return-return)	0.17
C.3.20 Changes in book equity	0.59	E.5.41 Earnings persistence	0.20
C.3.12 Composite equity issuance	0.67	F.6.3 Idiosyncratic volatility	0.23
		C.3.25 Net equity financing	0.24
		F.6.11 Share turnover	0.28
		E.5.46 Asset liquidity	0.30
		F.6.25 Liquidity betas (net)	0.32

Panel B: Max and Min of λ are 0.97 and 0.03, respectively

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Eigenvalues	16.81	2.61	0.65	0.79	0.53
Correlated Assets	202	169	149	136	90

Coefficients of candidate factors associated with the latent factors

Factor 2		Factor 3		Factor 4		Factor 5	
A.2.3 Quarterly book-to-market equity	-0.24	D.4.12 Operating profits to equity	-0.27	E.5.8 Operating leverage	-0.5	E.5.10 Hiring rate	-0.57
F.6.25 Liquidity betas (net)	-0.22	E.5.46 Asset liquidity	-0.27	F.6.15 Share price	-0.34	E.5.11 R&D capital-to-assets	-0.43
F.6.17. Turnover-adjusted number of zero daily volume	-0.13	F.6.12 Coefficient of variation of share turnover	-0.24	F.6.17 Turnover-adjusted number of zero daily volume	-0.32	E.5.4 R&D expense-to-market	-0.18
E. 5.51 Average returns Rn[2,5]	-0.07	F.6.25 Liquidity betas (net)	-0.16	F.6.17. Turnover-adjusted number of zero daily volume	-0.19	E.5.51 Average returns Rn1	-0.16
D.4.6 Capital turnover	-0.03	C.3.20 Changes in book equity	-0.11	F.6.1 Market equity	-0.18	D.4.15 Operating profits-to-assets	-0.07
E.5.46 Asset liquidity	-0.02	D.4.5 Profit margin	-0.06	D.4.32 Book leverage	-0.07	E.5.8 Operating leverage	-0.04
F.6.25 Liquidity betas illiquidity-illiquidity	-0.02	A.2.16 Net payout yield NOP	-0.03	C.3.20 Changes in book equity	-0.02	C.3.20 Changes in book equity	0.00
E.5.32 Financial constraints (the Whited-Wu index)	-0.01	D.4.16 Operating profits-to-lagged assets	0.04	E.5.30 Financial constraints (the Kaplan-Zingales index)	0.11	F.6.25 βret1, βret6, and βret12, liquidity betas (return-return)	0.00
D.4.7 Quarterly assets turnover	0.00	E.5.25 Labor force efficiency	0.06	E.5.4 R&D expense-to-market	0.12	E.5.51 Average returns Ra[6,10]	0.03
D.4.12 Operating profits to equity	0.00	E.5.51 Average returns Rn1	0.06	D.4.5 Assets turnover	0.13	C.3.25 Net equity financing	0.05

E.5.45 Pension plan funding rate	0.00	E.5.11 R&D capital-to-assets	0.07	F.6.8 Market beta	0.15	E.5.46 Asset liquidity	0.08
E.5.45 Pension plan funding rate	0.00	A.1.5. Prior 11-month returns	0.17	E.5.46 Asset liquidity	0.18	E.5.51 Average returns Rn[6,10]	0.12
E.5.51 Average returns Ra[6,10]	0.00	E. 5.51 Average returns Ra[2,5]	0.18	F.6.11 Share turnover	0.18	A.2.2 Book-to-June-end market equity	0.18
E.5.30 Financial constraints (the Kaplan-Zingales index)	0.06	F.6.6 Total volatility	0.19	F.6.25 Liquidity betas (return-return)	0.22	E.5.25 Labor force efficiency	0.60
E. 5.51 Average returns Ra[2,5]	0.08	E. 5.51 Average returns Rn[2,5]	0.35	F.6.3 Idiosyncratic volatility	0.36		
F.6.14 Coefficient of variation of dollar trading volume.	0.08	C.3.25 Net equity financing	0.42	F.6.25 Liquidity betas (net)	0.39		
F.6.25 Liquidity betas (return-return)	0.29	E.5.30 Financial constraints (the Kaplan-Zingales index)	0.58				
C.3.20 Changes in book equity	0.61						
C.3.12 Composite equity issuance	0.63						

Table 7 IPCA characteristic selection

We assume that the factors with time-varying factor loadings determine the testing asset. These loadings are the linear functions of the firm characteristics. The factors are unknown and should be estimated. The loadings are also unknown. The firm return and characteristics are known. We apply SOFAR to estimate the factors, the select number of factors, estimate coefficients of factor loadings on firm characteristics, and select the characteristics with the highest impact on factor loadings, based on Equation (**7**) in Section 2. The table shows the characteristics being selected. The number of stocks included in each year is presented in Panel A of Table 1. There are 84 characteristics, as in Panel B of Table 1We present the results in four sample periods: 1980-2018, 1970-2015, 1995-2018, and 1980-2003.

1980-2018		1970-2015		1995-2018		1980-2003	
beta	-0.35	beta	-0.19	beta	-0.13	beta	-0.27
mve	0.63	mve	0.62	mve	0.52	mve	0.56
baspread	0.00	mom12m	0.61	baspread	-0.32	baspread	-0.13
mom12m	0.59	mom1m	0.35	mom12m	0.39	mom12m	0.20
mom1m	0.35	mom36m	0.26	mom1m	0.36	mom1m	0.19
std_dolvol	0.12	zerotrade	0.13	sgr	0.16	std_dolvol	0.40
				pchsale_pchxsga	-0.24	zerotrade	0.59
				chnanalyst	0.36		
				chcsho	0.08		
				hire	-0.23		
				salecash	0.23		

Table 8 Out-of-sample predictions

There are five methods to be compared in the out-of-sample prediction: the RRA method with constraints on latent factors, asset, and candidate factor selection, the RRA together with selecting a pre-specified the number of latent factors (five factors in this case), the RRA with a constraint on latent factors, the RRA with constraints on latent factors and asset selection, and the RRA with constraints on latent factors and candidate factor selection. These methods are associated with equation (2) with all λ_d, λ_a and λ_b turned on, equation (3) with the first five latent factors, with equation (2) with only λ_d turned on, with equation (2) with λ_d and λ_b turned on, and with equation (2) with λ_d and λ_a turned on. The first 30 years are the training periods, and the remaining years are used for out-of-sample prediction. The sum of alpha squares and the total R-square over all 202 portfolios are defined following Huang, Li, and Zhou (2019). Panel A presents the results for 202 Giglio and Xiu portfolios and Panel B presents the results for 2190 deciles portfolios sorted by the characteristics from Hou, Xue and Zhang (2019).

Panel A

	Sum_alpha_2	**R_square**
All constraints	0.10	0.58
Five factors	0.28	0.55
Constraints on latent factors	0.36	0.31
Constraints on latent factors and asset selection	0.23	0.27
Constraints on latent factors and factor selection	0.07	0.59

Panel B

	Sum_alpha_2	**R_square**
All constraints	0.06	0.70
Five factors	0.13	0.63
Constraints on latent factors	0.28	0.38
Constraints on latent factors and asset selection	0.20	0.41
Constraints on latent factors and factor selection	0.09	0.71