Interpretable Characteristics-based Factors: A Machine Learning Approach

First Draft: July, 2020
Current version: January, 2023

Interpretable Characteristics-based Factors: A Machine Learning Approach

Abstract

We propose a new approach to construct factors from firm characteristics. In contrast to existing studies, each of our factors comes from the same group of statistically related firm characteristics, making its economic interpretation straightforward. The number of groups is not chosen ad hocly, but rather determined by data. Applying our method to a set of 94 representative firm characteristics, we find that the factors chosen by our approach is not only easy to interpret economically, but the associated factor model outperforms existing models, in particular improving the recent Instrumented Principal Components Analysis (IPCA) model of Kelly, Pruitt and Su et al. (2019) and related recent machine learning models. Further Bayesian model comparison reaffirms the conclusion.

JEL Classification: G11, G14

Keywords: Factor model, Cross-sectional stock return, Cluster analysis, Model comparison, IPCA, Neural networks

1. Introduction

As surveyed by Harvey, Liu, and Zhu (2016) and Hou, Xue, and Zhang (2020), there are potentially hundreds of firm characteristics or firm-level factors that affect the expected returns in the cross section of stocks. Following Cochrane (2011), there are two important questions. First, how many factors do we really need? Second, given a set of well-known factors, such as the prominent five factors of Fama and French (Fama and French, 2015), are there other factors that can provide incremental information for explaining the cross-sectional variation of expected stock returns?

There are mainly two existing approaches to answer the above questions. The first is primarily the principal component analysis (PCA). Early studies, such as Connor and Korajczyk (1986), apply it to extract factors from returns. Though PCA can be applied to a panel of firm characteristics easily, the problem is that the extracted factors are linear combinations of *all* the existing characteristics, making them difficult to interpret economically. For example, for a set of 4 characteristics with two of them being value and two being growth, any PCA factor will be a combination of the 4, resulting a neither value nor growth factor. Recently, Kelly et al. (2019) pioneer a new PCA method, the Instrumented PCA, which allows factor loadings to be characteristic-dependent. However, the resulted IPCA factors remain difficult to interpret. The second approach is the growing application of machine learning (ML) to finance. Feng, Giglio, and Xiu (2020) and Freyberger, Neuhierl, and Weber (2020), Han et al. (2021) and Kozak, Nagel, and Santosh (2020), among others,[1] propose various ML methods to identify which firm characteristics drive the stock returns. However, this literature tends to over-identify the number of factors that matter, as it is difficult to handle and distinguish highly correlated factors by the existing ML models.

Our paper provides a simple approach to adress the problem. Intuitively, our method

[1]Other examples include Cong et al. (2021), DeMiguel et al. (2020), Gu, Kelly, and Xiu (2020), Daniel et al.(2020), Chen and Velikov (2020), Chordia, Goyal, and Saretto (2020),Patton and Weller (2020), and Avramov, Cheng, and Metzker (2021).

has two steps. First, we divide the factors into statistically related clusters. If the factors have the same economic source, they must be statistically related. In the second step, we extract optimally a factor for each cluster that capture better the economic driving force. Our approach is inspired by Stambaugh and Yuan (2017) who appear the first to use cluster to isolate the factors. In contrast to their method, we apply a clustering algorithm that is applicable to high dimensional case, and can determine the number of clusters from data without imposing it a priori. In addition, instead of taking the average of the factors in a cluster as the new factor, we let data to find the best factor from the cluster.

We apply our method to a set of 94 representative firm characteristics as used by Gu et al. (2020), which is similar to the data sets used by many others. In terms of using clustering algorithms, our paper appears the first to examine such a large data set in finance. Furthermore, we find that the resulted factor models outperform well known models, including the IPCA of Kelly, Pruitt and Su et al. (2019). Moreover, the ML literature finds that there are more than 20 characteristics that are critical in determining the variation of cross-sectional expected returns. We find that there are only 9 clusters that imply 9 factors. The resulted long-short portfolios also outperform those of the recent ML models.

We also compare models from a Bayesian perspective. Given the large number of possible combinations of factors in a model, standard econometric techniques, mainly developed for evaluating the adequacy of a single model, is not sufficient for identifying the best factor pricing model(s). As argued by Lee and Potscher (2005), due to model uncertainty, models selected from different target functions do not converge asymptotically to the same limit. Barillas and Shanken (2018), along with Chib et al. (2020) (henceforth BS-CZZ), develop a general Bayesian model comparison method invariant to test portfolios in testing factor models. This method is similar but better than BIC coefficient, a widely used model comparison method in machine learning, which compares models based on their maximum likelihood estimations. In contrast, the BS-CZZ method compares models based on a wide range of model parameters.

We find that our model outperforms benchmark models including a model based on clustering factors according to economic concept, FF3, Q4, FF5 and Carhart model. Our analysis is carried out with various model comparison measures, including out-of-sample Sharpe ratio of the maximal Sharpe ratio portfolio, alpha test and BS-CZZ method.

The paper is organized as follows. Section 2 describes the data and cluster method we use. Section 3 describes the way to construct factor models based on cluster analysis. Section 4 presents the empirical results on performance of the factor models. Section 5 presents the empirical results on that the cluster method improves the IPCA and ML models. Section 6 presents the robustness results. Section 7 concludes.

2. Data and Cluster Method

2.1. Data

We use the 94 characteristics used in Gu et al. (2020) in the US market. Firm characteristics data is from https://dachxiu.chicagobooth.edu/. Detailed characteristics definition can be found in Table A.6 of Gu et al. (2020). Monthly stock return data is from CRSP.

The data clear process is similar to Gu et al. (2020). Firstly, we replace missing firm characteristics with cross-sectional median at each month if the number of non-missing observations at that month is more than 500. The purpose of replacing missing values is to avoid discarding an observation only because one of firm characteristics is missing. Secondly, according to Gu et al. (2020), we include stocks with prices below $5, share codes beyond 10 and 11, and financial firms. Finally, we only keep stocks listed on three major exchanges, NYSE, AMEX and NASDAQ. Data starts from January 1985 and ends in December 2021.

2.2. Motivation for using cluster method

Cluster analysis groups a set of objects into subsets or "clusters" such that those within each cluster are more closely related to one another than objects assigned to different clusters (e.g., Hastie et al., 2009). Fundamental to cluster analysis is the definition of similarity or proximity between objects. This can only come from subject matter considerations and is similar to the specification of a loss function in prediction problems (supervised learning). In our study, we apply cluster analysis to the firm characteristics. Following Stambaugh and Yuan (2017), a natural definition of the similarity between the firm characteristics can be captured by the cross-sectional correlations between them.

Previous researches construct tradable factor portfolios based on Fama-Macbeth regression (Lewellen, 2015; Han et al., 2021). Fama-Macbeth regression is first proposed by Fama and Macbeth (1973) to examine relationship between return and risk:

$$r_{t+1} = \beta X_t + e_t, \tag{1}$$

where $X_t \in R^{N\times I}$ is I firm characteristics of N firms in month t, and $r_{t+1} \in R^N$ is the excess returns of N assets in month $t+1$. β ,the relationship between return and risk, is estimated with average of cross-sectional regression coefficients

$$\hat{\beta} = \frac{1}{T}\sum_{t=1}^{T} \hat{\beta}_{t+1}, \tag{2}$$

where

$$\hat{\beta}_{t+1} = (X_t^T X_t)^{-1} X_t^T r_{t+1}. \tag{3}$$

Barra (1998) finds that the estimator $\hat{\beta}_{t+1}$ are returns of tradable factor portfolios with stock weights

$$w = (X_t^T X_t)^{-1} X_t^T. \tag{4}$$

Lewellen (2015) and Han et al. (2021) construct tradable factor portfolios based on an out-of-sample estimate of β with rolling average of $\hat{\beta}_{t+1}$.

However, previous literature, such as Lewellen (2015) and Han et al. (2021), did not construct factor models with Fama-Macbeth regression. Motivated by their method, we construct factor models with Fama-Macbeth regression after the cluster analysis. We assume that the true model consists of a few latent factors, firm's exposures to which are observable through firm characteristics with measurement error. We can filter out the error by grouping together the firm characteristics that indicate the same signal. Specifically, for each time $t = 1, ..., T$, let $X_{it}, \{i = 1, ..., I\}$ be I firm characteristics, with each $X_{it} \in R^N$. Assume K factors $F_{kt}, \{k = 1, ..., K\}$, each observable through several firm characteristics with error. Let $\mathcal{P}_k, \{k = 1, ..., K\}$ be a partition of the integer set $\{1, ..., I\}$, then each X_{it} is associated with one F_{kt} if $i \in \mathcal{P}_k$, i.e.,

$$X_{it} = F_{kt} + e_{it}, \ i \in \mathcal{P}_k, \tag{5}$$

where $e_{it} \in R^N$ are independent error terms and homogeneous within each cluster. For firm characteristics in each group $\mathcal{P}_k$, we can construct a factor using Fama-Macbeth regression. Thus, with K groups, we get a K-factor model. Details on factor model construction method is in section 3.1.

Generally, cluster algorithms fall into three distinct types: combinatorial algorithms (e.g. hierarchical clustering), density-based algorithms (e.g. PRIM) and spectral clustering (Hastie et al., 2009). We use a form of combinatorial algorithm, the hierarchical clustering analysis (HCA), because it works directly on the observed data with no direct reference to an underlying probability model. Spectral clustering also does not assume any probability model. We do not use it because it is based on principal component analysis and sensitive to noise in data (Bojchevski et al., 2017), while financial data is notorious for its low signal-to-noise ratio. We overcome the problem of low signal-to-noise ratio with a hierarchical clustering algorithm called Chameleon (Karypis et al., 1999), which is suitable for the

data set with arbitrary density. Specifically, if a firm characteristic measures a risk exposure with large measurement error, it will be dissimilar to other firm characteristics. Chameleon will divide firm characteristics with above features in a cluster with low density rather than divide each of them as a cluster. Some modified versions based on Chameleon appear after the publication of this algorithm, but their modifications are small and Chameleon remains the most popular version among them. Implementation details of Chameleon are presented in the next subsection.

2.3. Clustering firm characteristics

In this section, we introduce our cluster method in detail. First, we introduce a concept of Intuitive Cluster (IC), which is used as prior information in our cluster method. Then, we define similarity between firm characteristics. Finally, we show steps of Chameleon to divide firm characteristics into clusters.

IC is an existing ad hac way to divide firm characteristics. It divides firm characteristics into 6 clusters based on the economic concept and is widely used in the literature (Hou et al., 2015; Hou et al., 2020; Han, et al., 2020).

Consistent with Hastie et al. (2009) and Stambaugh et al. (2017), we define the similarity as the absolute value of time-averaged value-weighted cross-sectional Spearman correlations. Different from Stambaugh et al. (2017), we use value-weighted correlations to put more emphasis on large valued stocks. Let x_{it}^n to be the cross-sectional rank of the firm characteristics X_{it}^n, the time averaged weighted rank correlation is

$$\rho_{ij} = \left| \frac{1}{T} \sum_{t=1}^{T} \frac{\sum_{n=1}^{N_t} w_t^n (x_{it}^n - \bar{x}_{it})(x_{jt}^n - \bar{x}_{jt})}{\sqrt{\sum_{n=1}^{N_t} w_t^n (x_{it}^n - \bar{x}_{it})^2} \sqrt{\sum_{n=1}^{N} w_t^n (x_{jt}^n - \bar{x}_{jt})^2}} \right|, \tag{6}$$

where $\bar{x}_{it} = \sum_{n=1}^{N_t} w_t^n x_{it}^n$, w_t^n is the value weight parameter for each stock n and $\sum_{n=1}^{N_t} w_t^n = 1$.

In Karypis et al. (1999), the clustering problem is represented by a graph, where, in our case, graph vertices represent firm characteristics, and weighted edges represent similarities among the firm characteristics. The main feature of the clustering algorithm by Karypis et al. (1999) is to use two pass HCA with the first pass the divisive HCA and the second pass the agglomerative HCA (e.g. Hastie et al. 2009). In the first pass, the algorithm applies a graph partitioning method to find the min-cut partition of clusters with roughly equal size (Karypis et al. 1999). In the second pass, the algorithm proceeds sequentially by optimally merging clusters. For two clusters C_i and C_j , the algorithm seeks to minimize inter-cluster similarity, which is customarily defined suitable for specific problem. In our paper, we seek to minimize the following:

$$IS(C_i, C_j) = \frac{\bar{S}_{EC_{\{C_i,C_j\}}}}{\frac{|C_i|}{|C_i|+|C_j|}\bar{S}_{EC_{C_i}} + \frac{|C_j|}{|C_i|+|C_j|}\bar{S}_{EC_{C_j}}}, \tag{7}$$

where $\bar{S}_{EC_{\{C_i,C_j\}}}$ is inter-cluster closeness, measured as average weights of the edges that connect vertices in C_i to vertices in C_j. $\bar{S}_{EC_{C_i}}$ is within-cluster closeness, measured as the average weight of the edges that belong in the min-cut bisector of cluster C_i. $|C_i|$ is number of vertices in C_i. Intuitively, the inter-cluster similarity so defined captures the relative similarity between clusters compared to the intra-cluster similarity.

An overview of the method is provided by Figure 1 (copied from Karypis et al., 1999). Specifically, the algorithm is implemented as follows.

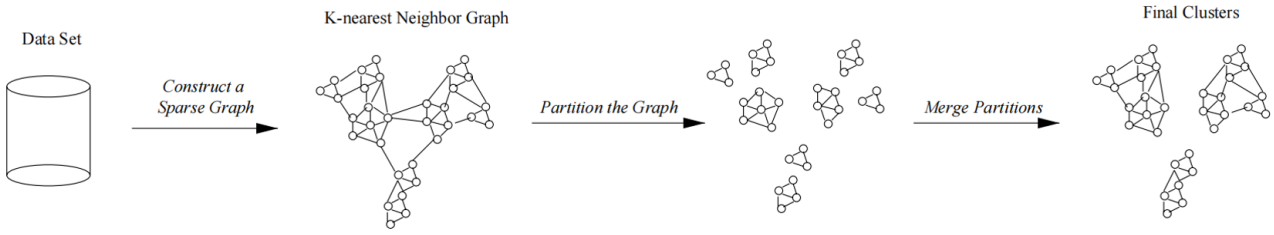


Figure 1: Clustering process

First, a sparse graph is obtained through nearest neighbor method, i.e. the edge between two vertices will be kept only if either of the graph vertices is among the set of knn (a hyperparameter of the model) nearest neighbors of the other vertice, otherwise the weight of the edge will be set to zero. This is a conventional step for graph clustering because it helps to reduce noise.

Second, in the divisive HCA, the algorithm repeatedly applies a graph partitioning algorithm until graph vertices are divided into m clusters. The graph partitioning algorithm finds the partition of a given graph with minimum edge-cut, defined as the sum of the weight of the edges that straddle partitions (refer to Karypis et al. 1999 for details). The purpose of this step is to divide the graph into very small groups, each consisting of 3 or 4 vertices. These groups are used to initialize the intra-cluster similarity. In our paper, it is this step that we impose the prior information of IC. That is, we guarantee that the vertices (firm characteristics) in the same cluster also belong to the same cluster in IC by replacing the similarity between firm characteristics in different IC clusters to be 0. This restriction is lifted in later steps.

Thirdly, a version of agglomerative HCA algorithm is applied and clustering results are obtained for all possible numbers of clusters K. In this step, we merge clusters with the highest inter-cluster similarity defined in equation (7).

So far, we have described the mechanism of the algorithm of Karypis et al. (1999) and our modifications. There are three hyper-parameters $\{knn, m, K\}$ to be determined, where knn is the number of the nearest neighbors, m is the number of clusters after the first pass, and K is the number of clusters after the second pass. We select hyper-parameters from a grid of values: $knn = \{10, 15\}, m = \{24, 31\}$ and $K = \{1, 2, ..., 15\}$. m is chosen such that 3 or 4 vertices are in each cluster after the first pass.

To determine hyper-parameters, We define the performance measure as the average of

inter-cluster similarity , expressed as

$$\overline{IS} = \frac{2}{K(K-1)} \sum_{i,j=1,\ldots,K, i \neq j} IS(C_i, C_j), \tag{8}$$

where $IS(C_i, C_j)$ is defined in equation (7) and K is the number of clusters. We choose the clustering result with the smallest $\overline{IS}$ and the corresponding hyper-parameters are $knn^* = 15, m^* = 31$ and $K^* = 9$.

Figure 2 demonstrates the performance measure $\overline{IS}$ versus the number of clusters K given the optimal hyper-parameters $knn^* = 15$ and $m^* = 31$. The curve is approximately of "U" shape. $\overline{IS}$ first decreases and then increases as the number of clusters increases. The minimum $\overline{IS}$ is obtained at $K = 9$. In other words, at the beginning of the agglomerative HCA, merging two clusters tends to reduce $\overline{IS}$. However, after we have gotten 9 clusters, further merging tends to enlarge $\overline{IS}$. According to equation (7), it means further merging tends to generate higher between-cluster similarity ($\bar{S}_{EC_{C_i}}$), or lower intra-cluster similarity ($\bar{S}_{EC_{C_i}}$), either of which indicates that we should stop merging. Thus, we choose the result with 9 clusters.

2.4. Data-driven Clustering vs Intuitive Clustering

In this section, we present difference between our clustering result (Data-driven Clustering, denoted as DC) and IC.

In Table 1, we use identifiers that from DC1 to DC9 to denote 9 clusters in DC, and identifiers that from IC1 to IC6 to denote 6 clusters in IC. Panel A presents the number of firm characteristics in each cluster of DC and IC. It shows that the cluster *Trading frictions* in IC is splitted into four clusters in DC, which are *Illiquidity*, *Trading frictions* (*measured by volume*), *Trading frictions* (*measured by return*) and *Beta*, *resp.* . *Illiquidity* includes Amihud illiquidity, size, etc., *Trading frictions* (*measured by volume*) includes characteristics con-

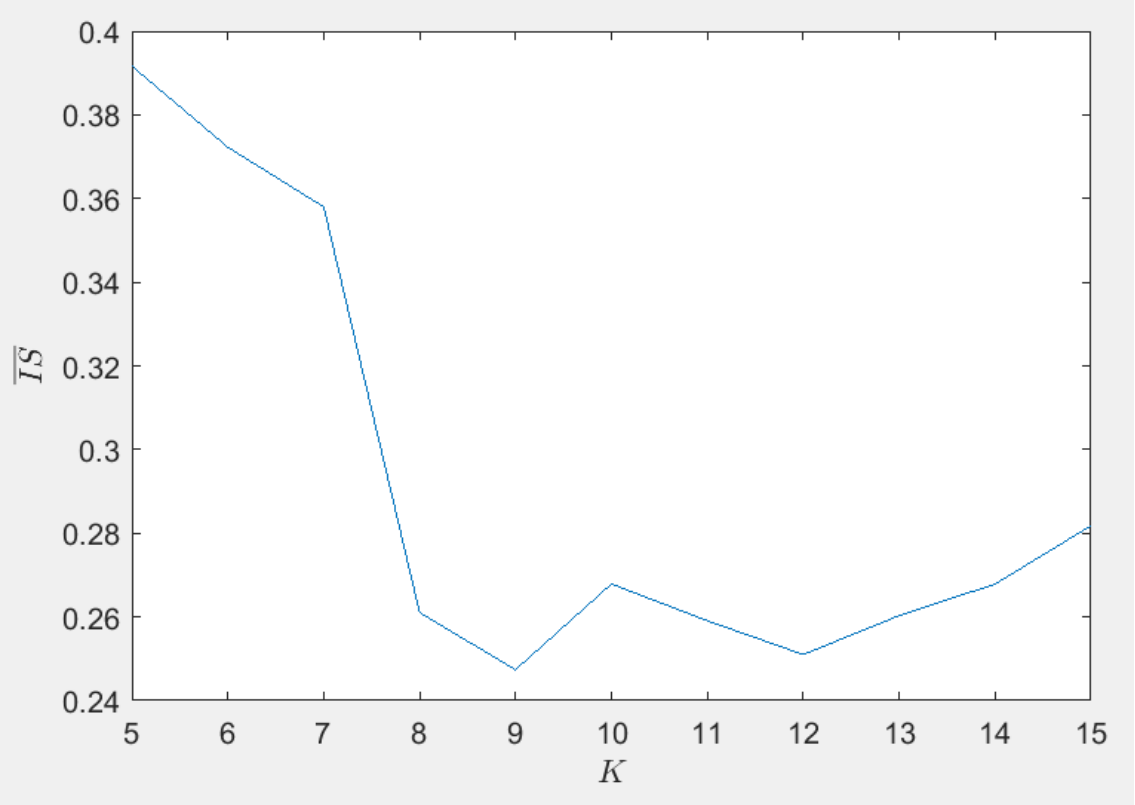


Figure 2: Average inter-cluster similarity
The figure presents the relationship between average inter-cluster similarity $\overline{IS}$ in equation (8) and the number of clusters K. The horizontal line is the number of clusters K and the vertical line is average inter-cluster similarity $\overline{IS}$.

structed based on trading volume, such as share turnover. $Trading\ frictions$ ($measured\ by\ return$) includes characteristics constructed based on trading price, such as idiosyncratic return volatility, maximal return and so on. $Beta$ contains beta and beta square.

Additionally, firm characteristics in $Momentum$ in IC are splitted into three clusters in DC. The first one is $Long-run\ momentum$, including 12-month momentum, 6-month momentum. The second one is $Short-run\ momentum$, including 1-month momentum, etc.. The third one is $Growth$.

Moreover, firm characteristics in $Profitability$, $Value$, $Investment$ and $Intangible$ in IC are re-clustered into $Accruals$, $Profitability$ and $Growth$ in DC.

We will construct factor models based on DC and demonstrate more properties of our

method in the next section.

3. Clustered Factor Models (DC-Model)

In this section, we first construct factor models based on DC and IC, and show performnace of factors in those models.

3.1. Construction of factor models through 3-step regression

As demonstrated in the previous section, given a clustering result $\mathcal{P}_k, \{k = 1, ..., K\}$, we use Fama-Macbeth regression to construct a factor model in 3 steps.

First, for each $k = 1, 2, ..., K$ and in each month t, we run a cross-sectional OLS regression:

$$R_t = \hat{\beta}_{0,k,t} + \sum_{i \in P_k} \hat{\beta}_{i,k,t} \hat{X}_{i,t-1} + \epsilon_t, \tag{9}$$

where R_t is n-dimensional stock returns in month t, and $\hat{X}^n_{i,t-1}$ is the i^{th} standardized firm characteristics in month $t-1$. In the appendix, we prove that under certain technical conditions, the predictive regression of (9) converges to the true factor model.

Second, the predicted returns with firm characteristics in cluster P_k in month t is

$$\hat{R}^n_{k,t+1} = \bar{\beta}_{0,k} + \sum_{i \in P_k} \bar{\beta}_{i,k} \hat{X}^n_{i,t}, \tag{10}$$

where $\bar{\beta}_{i,k,t} = \frac{1}{t} \sum_{\tau=1}^{t} (\hat{\beta}_{i,k,\tau})$, is the smoothed OLS (SOLS) estimator as in Han et al. (2020), with $\hat{\beta}_{i,k,\tau}$ the OLS estimator in the cross-sectional regression (9). Then we use portfolio sorting method based on the SOLS predictor to construct corresponding factor portfolios. In particular, we rank the predicted return $\hat{R}^n_{k,t}$ in (10) from high to low. We define the factor portfolio $Y_{k,t}, (k = 1, 2, ..., K)$ to be a long-short zero investment portfolio that long

the upper median and short the lower median. Formally, the weight of stock n in the factor portfolio $Y_{k,t}$ formed in the end of month t is defined as

$$Y_{k,t}^n = \begin{cases} 2/N, & \hat{R}_{k,t+1}^n \text{ is in top } 50\%; \\ -2/N, & \hat{R}_{k,t+1}^n \text{ is in bottom } 50\%. \end{cases} \tag{11}$$

Third, we obtain K independent factors as model factors based on regression (12)

$$R_{t+1}^n = \gamma_0 + \sum_{k=1}^{K} \gamma_{k,t} Y_{k,t}^n + e_{t+1}^n, \tag{12}$$

where the stock weight $Y_{k,t}^n$ in the factor portfolio $Y_{k,t}$ is used as a measure of risk exposure of stock n on the k^{th} factor and $\gamma_{k,t}$ measures the risk premium of the k^{th} factor.

Weights of the k^{th} independent factor portfolio is given by the k^{th} column vector of ω_t as

$$\omega_t = W_t Y_t (Y_t' W_t Y_t)^{-1}, \tag{13}$$

where Y_t is the matrix consisting of column vector $Y_{k,t}$. We scale $Y_{k,t}$ to make ω_t have a \$1-long-position. For each factor portfolio $Y_{k,t}$, the corresponding risk premium is $\gamma_{k,t}$, the regression coefficient of (12), where W_t is the weighting matrix for the regression at time t. We use value-weighted diagonal matrix for W_t.

So far we have used a 3-step regression to construct the DC-model with K factors. In the next subsection, we compare the performance of DC-Model and corresponding IC-Model, the so constructed factor model based on IC.

3.2. Performance of factors in DC-Model and IC-Model

This section presents performance of factors in DC-Model and IC-Model. Table 2 presents the result. Panel A reports performance of Trading frictions factors, 1 in IC-Model and 4

in DC-Model. Panel B reports performance of Momentum factors, 1 in IC-Model and 2 in DC-Model. Panel C reports performance of other factors. It is shown that by splitting $Trading\ frictions$ cluster in IC, we get factors with higher Sharpe ratios. $Trading\ frictions$ in IC are splitted into 4 clusters in DC. Corresponding factors of those 4 clusters achieve Sharpe ratios of 0.32, 0.34, 0.11 and 0.16, all of which are much higher than 0.07, Sharpe ratio of trading frictions factor in IC-Model. Besides, by splitting $Momentum$ in IC, we also obtain a factor with higher Sharpe ratio. The long-run momentum factor in DC-Model achieves Sharpe ratio of 0.83, which is higher than 0.42, Sharpe ratio achieved by momentum factor in IC-Model. The short-run momentum factor in DC-Model achieves Sharpe ratio of 0.36, which is comparable to that of momentum factor in IC-Model (0.42).

Table 3 presents correlations of factors in DC-Model and IC-Model and there are two findings. First, factors in DC-Model splitted from IC are not highly correlated. For example, the $Trading\ frictions$ in IC (IC1) is splitted into four clusters by DC (DC1-DC4). Factors corresponding to DC1-DC4 are not highly correlated, with the highest correlation of 0.44 coming from that between DC2 and DC3. Second, table 3 shows that by splitting a cluster in IC, the resulting factors in DC-Model are not all highly correlated with the factor in IC-Model. For example, factors corresponding to DC1 is not correlated with the factor corresponding to IC1. It means that by splitting $Trading\ frictions$ in IC, we get a factor with additional information.

Results in Table 2 demonstrate that our method contributes to extract information more efficiently than IC. We further illustrate that our so constructed DC-Model outperforms popular benchmark models as well in the next section.

4. Model Comparison Tests

In this section we perform various model comparison tests between DC-Model and bench-

mark models. We use three methods to compare models, construction of out-of-sample (OS) maximal Sharpe ratio portfolios, alpha test and a Bayesian method. The benchmarks are IC-Model, FF3, Car4, Q4 and FF5. FF3 is first proposed by Fama and French (1993). They use size and BM to construct factors. Those two factors and market factors constitute FF3 factor model. Carhart (1997) adds momentum factor and constitute Car4 factor model. Hou, Xue and Zhang (2015) propose Q4 model, where factors are constructed on size, ROE, investment and market. Fama and French (2015) modify FF3 and add operating profitability factor and investment factor, proposing the FF5 factor model.

4.1. Performance of OS maximal Sharpe ratio portfolio

In this section we introduce the method to form an OS maximal Sharpe ratio portfolio and present its empirical results. This test is equivalent to searching a Stochastic Discount Factor (SDF) close to the Hansen-Jagannathan bound (Hansen and Jagannathan, 1991). Since the SDF is constructed by OS tradable portfolios, higher OS maximal Sharpe Ratio implies better spanned pricing model.

Given K factors $F_t = (F_{1,t}, F_{2,t}, ..., F_{K,t})'$, we can write the SDF, denoted by M_t, as

$$M_t = 1 - b'(F_t - EF_t), \tag{14}$$

where b is a $K \times 1$ vector of constant. The unconditional asset pricing implies that M_t satisfies the following equation,

$$E[M_t F_t] = 0. \tag{15}$$

Substituting Equation (14) into (15), we obtain

$$b = \Sigma^{-1} E(F_t), \tag{16}$$

where Σ is the covariance matrix of factors.

Assume the estimated mean and covariance matrix of factors are denoted as $\bar{\mu}$ and $\bar{\Sigma}$, the estimator for b in Equation (16) is

$$\hat{b} = \bar{\Sigma}^{-1}\bar{\mu}, \tag{17}$$

and the variance of SDF can be written as

$$\sigma^2(M_k) = \bar{\mu}^T\bar{\Sigma}^{-1}\bar{\mu}. \tag{18}$$

When using the sample moment conditions, $\sigma(M_k)$ is also known as Hansen-Jagannathon bound, the upper bound for Sharpe ratios of all investable portfolios. Note that this upper bound is not achievable by investable portfolios unless the parameters are constant.

To achieve the maximal investable Sharpe ratio, we use the conditional version of

$$E_t[M_{t+1}F_{t+1}] = 0. \tag{19}$$

That is, we use rolling sample estimation for mean and variance μ_t and Σ_t. At time t, we hold the optimal portfolio consisting of K factor portfolios as

$$b_t = \Sigma_t^{-1}\mu_t. \tag{20}$$

Note that when b_t is computed using out-of-sample Σ_t and μ_t, the portfolio is investable, and it is the out-of-sample efficient frontier portfolio given the factor model. We call the Sharpe ratio of this portfolio the out-of-sample (OS) maximal Sharpe ratio of the factor model and call this portfolio the out-of-sample (OS) maximal Sharpe ratio portfolio of the factor model. The return of an OS maximal Sharpe ratio portfolio is constructed on a purely out-of-sample basis by using the mean and covariance matrix of estimated factors through t and tracking the post-formation $t+1$ return.

In addition to factor models mentioned in the beginning of section 4, we also compare with some machine learning approaches, such as Lasso, Ridge and Enet. We obtain return predictions of those methods with panel regression in a rolling window and construct long-short zero-investment portfolios. Specifically, at the end of each month t, we predict returns in month $t+1$ with data through month t, where hyper-parameters are chosen with cross-validation. Then we construct the portfolio where 50% stocks with the highest return predictions are in the long position and the rest are in the short position.

Recall that our data is from January 1985 to December 2021. Factors are from January 1987 to December 2021, since we skip the first 24-month data to estimate predicted returns in step (10). In addition, the OS maximal Sharpe ratio portfolios are from January 1990 to December 2021, since we skip the first 60-month data to calculate conditional mean and covariance.

Table 4 shows performance of the OS maximal Sharpe ratio portfolios for DC-Model and benchmarks. It is shown that for all measures DC-Model outperforms benchmarks. The average monthly return of 0.53% for DC-model is the highest, and is about 0.06% higher than the next best value of 0.47%, achieved by Ridge regression. As for standard deviation, DC-Model has the lowest value of 1.38%, smaller than the next best value of 1.43%, achieved by IC-Model. The Sharpe ratio and maximal drawdown (MDD) deliver perhaps the sharpest differences between DC-Model and benckmarks. DC-Model has Sharpe ratio of 1.34 and MDD of 8.22%. In contrast, Sharpe ratios are only 0.91 for FF5 and 0.9 for IC-Model, and MDDs are larger than 17% for all benchmarks.

4.2. Alpha test

In this section, we introduce implementation and empirical results of alpha test. Assume we want to compare models $M_1, M_2, ..., M_L$, we run the regression

$$f_t = \alpha_l + \beta_l F_{l,t} + e_{l,t},\ l = 1, 2, ..., L, \tag{21}$$

where $f_t \in R^N$ are N test portfolios' returns in month t, $F_{l,t} \in R^K$ are K factors in M_l in month t.

Table 5 compares the models on several measures that summarize abilities to accomodate test portfolios' returns: the average absolute alpha, the number of anomalies for which the model produces the smallest absolute alpha among L models being compared, the number of anomalies for which p-value is smaller than 0.1, and 0.05. Table 5 also calculates measures based on adjusted p-value: the number of anomalies for which adjusted p-value is smaller than 0.1, 0.05, where the adjusted p-value is adjusted for 5-lags auto-correlation in error term $\{e_{l,t}\}_{t=1}^T$. Panel A reports these measures for the set of 94 anomalies. For each measure, we see the DC model performs the best, followed by the IC model. THe average absolute alpha of 0.08% for the DC models is about 0.02$ lower than the next best value of 0.10%, achieved by the IC model. For 27 of the 94 anomalies, the DC model achieves the lowest absolute alpha, compared to 22 anomalies for the IC model. Among 94 anomalies, 17 and 5 anomalies have alpha with p-value smaller than 0.1 and 0.05 for the DC model, compared to 22 and 13 anomalies for the IC model. The number of anomalies for which adjusted p-value is smaller than 0.1 and 0.05 follows a similar pattern.

Panel B report the same measure as panel A but for a different set of anomalies. we use 2 sets of test portfolios. Following Stambaugh et al. (2017), We reduce the set of 94 anomaly portfolios by excluding those most highly correlated with the factors in DC and IC models. This procedure leaves 47 anomalies. Panel B deliver essentially the same message

as those in panel A. In panel B, the gap becomes narrow for the number of anomalies for which p-value is smaller than 0.05 and adjusted p-avlue and is smaller than 0.1.

4.3. Bayesian model comparison

In this subsection, we compare models with a Bayesian method, first proposed by Barillas and Shanken (2018) and revised by Chib, Zeng and Zhao (2020) and thus is called BS-CZZ method hereafter. It is a more comprehensive test than alpha test. Alpha test has disadvantages in practice. As Barillas and Shanken (2018) said, "Although tests of the individual models are routinely reported, these tests often suggest 'rejection' of the implied restrictions, especially when the data sets are large (e.g., Fama and French, 2016). However, a relatively large p-value may say more about imprecision in estimating a particular model's alphas than the adequacy of that model."

Barillas and Shanken (2018) propose a Bayesian method to compute the posterior probability that a model holds conditional on data. The posterior probability that a model M_l holds, $Pr(M_l|D)$, is determined by marginal likelihood of two regressions (see equation (19) in Barillas and Shanken (2018))

$$f_{l,t} = \alpha_l + \beta_l Mkt_t + e_{l,t}, \tag{22}$$

$$f^*_{l,t} = \gamma_l f_{l,t} + \gamma_l^{Mkt} Mkt + u_{l,t}, \tag{23}$$

where $f_{l,t}$ is a vector that contains factors in model M_l except market factor in month t, $f^*_{l,t}$ is a vector that contains factors in all models except M_l in month t, and Mkt_t is market factor in month t. Marginal likelihood of equation (22) and (23) are called unrestricted and restricted marginal likelihood, denoted as $Pr_{U,l}$ and $Pr_{R,l}$, respectively.

For unrestricted regression (22), a key assumption is the prior for α_l. It is assumed that $Pr(\alpha_l|\beta_l, \Sigma_l) = MVN(0, k_l\Sigma_l)$, where Σ_l is covariance matrix of $e_{l,t}$ and k_l is prior assumed

ratio. k_l is informative about expected maximal Sharpe ratio that can be produced by combining factors in M_l, denoted as $E_l^{prior}(S_{max})$. Specifically, $k = \frac{(\kappa^2-1)S_{Mkt}^2}{K_l}$, where K_l is the dimension of f_t, S_{Mkt} is the (observed) Sharpe ratio of the market and

$$\kappa = E_l^{prior}(S_{max})/S_{Mkt}. \tag{24}$$

Table 6 applies the BS-CZZ method and compares DC-Model with benchmarks. It reports posterior probability $Pr(M_l|D)$ given prior for α_l expressed as κ in Equation (24). Panel A reports results of the comparison between DC-Model and each one of benchmark models by showing the posterior probability that DC-Model holds ($Pr(M_{DC}|D)$). We estimate κ as the ratio of equation (18) to the Sharpe ratio of the market portfolio. The former measures the expected maximal Sharpe ratio that can be produced by combining factors. As the estimated κ is about 3, we take κ around 3, that is, from 2 to 4. It is shown that under all κ, the probability is nearly 1.0, which means that the data strongly favors DC-Model. Panel B reports results of comparison among all models simultaneously and reports posterior probability that each model holds ($Pr(M_l|D)$). Results show a similar pattern as that in panel A. Under all κ, the probability that DC-Model holds is nearly 1.0, which means that the DC-Model is strongly favored to all benchmarks.

4.4. Analysis on out-of-sample clustering

So far the performance of DC-Model has been based on IS clustering. That is, DC is the clustering result with data in the full sample. Next, we analyze the performance of a factor model that is based on OS clustering.

To get an OS clustering result, we divide the 37 years of data into 27 years of training sample (1985-2011) and 10 years of testing sample. We get the OS data-driven clustering result (OSDC) with data in the training sample and examine performance of the correspond-

ing factor model (OSDC-Model) in the testing sample. Table 7 shows performance of the OS maximal Sharpe ratio portfolios in the testing sample. Generally speaking, OSDC-Model outperforms benchmarks. The average monthly return of 0.51% for OSDC-model is higher than IC-Model, Car4 and FF5 models, Lasso, Ridge and Enet methods, but is lower than FF3 and Q4 models. As for standard deviation, OSDC-Model has the lowest value of 1.36%, smaller than the next best value of 1.47%, achieved by IC-Model. The Sharpe ratio delivers perhaps the sharpest differences between OSDC-Model and benckmarks. OSDC-Model has the Sharpe ratio of 1.30 In contrast, Sharpe ratios are only 0.97 for Q4 and 0.92 for FF5. Besides, OSDC-Model has the lowest maximal drawdown (MDD). The MDD for DC-Model is 9.29%, smaller than the next best value of 10.28%, achieved by IC-Model.

5. Further Applications of Cluster Analysis

In previous sections, we construct DC-Model with Fama-Macbeth regression and demonstrate that the cluster analysis is useful to extract information from a large set of firm characteristics. In this section, we demonstrate that the cluster analysis can be complementary to existing IPCA method by Kelly et al. (2019) and neural networks.

5.1. Clustered factor model through IPCA

In this subsection, we construct a factor model through IPCA with complementary information from DC. IPCA is a Principal Component Analysis for conditional factor models proposed by Kelly et al. (2019), which further allows for time-varying risk exposures as linear functions of firm characteristics. Specifically, let

$$r_t = \beta_{t-1} f_t + e_t, \tag{25}$$

and

$$\beta_{t-1} = 1_N \Gamma_0 + X_{t-1}\Gamma + u_{t-1}, \tag{26}$$

where $r_t \in R^N$ consists of returns of N stocks in month t. $f_t \in R^J$ consists of J latent factors. $\beta_{t-1} \in R^{N\times J}$ consists of risk exposures of N stocks on J latent factors. $X_{t-1} \in R^{N\times I}$ consists of I firm characteristics of N stocks in month $t-1$. $\Gamma \in R^{I\times J}$ contains loadings of I firm characteristics on J risk exposures. Its element in the i^{th} row and the j^{th} column, Γ_{ij}, reflects the impact of i^{th} firm characteristics on j^{th} risk exposure. $1_N \in R^N$ is an all-one vector and $\Gamma_0 \in R^{1\times J}$ contains loadings of constant on J risk exposures. Its j^{th} element Γ_{0j} reflects average of j^{th} risk exposure across all stocks and months.

Given DC, we impose two restrictions in Equation (26). First, the k^{th} exposure is a linear function of firm characteristics in the k^{th} cluster, while loadings on other firm characteristics and constant are zero.

$$\Gamma_{ij} = 0, \text{ if } i \notin P_k \text{ and } j = 1, ..., K. \tag{27}$$

In addition, we add one more exposure that is irrelevant to all firm characteristics.

$$\Gamma_{i,K+1} = 0, \text{ for } i = 1, ..., I. \tag{28}$$

Next we construct IPCA model based on Equations (25) and (26), and IPCA with DC (denoted as IPCA+DC model) based on additional restriction in Equation (27) and (28). Estimation of IPCA model is based on value-weighted mean squared error (MSE), the first order conditions of which are given as

$$\hat{f}_t = (\hat{\Gamma}' X_{t-1}' W_{t-1} X_{t-1} \hat{\Gamma})^{-1} \hat{\Gamma}' X_{t-1}' W_{t-1} r_t, \tag{29}$$

and

$$vec(\hat{\Gamma}') = (\sum_{t=2}^{T} X_{t-1}' W_{t-1} X_{t-1} \otimes \hat{f}_t \hat{f}_t')^{-1} (\sum_{t=2}^{T} [X_{t-1} W_{t-1} \otimes \hat{f}_t]' r_t), \tag{30}$$

which are solved recursively as in Kelly et al. (2019). The IPCA+DC model is solved similarly with additional restrictions in Equation (27) and (28).

To make the estimation out-of-sample, at the end of each year, we use all data through the end of the year to solve equations (29) and (30) with or without restrictions of Equation (27) and (28) and get estimates of $\hat{\Gamma}$ for both models. Then we calculate the OS realized factor returns in each month of the following year with Equation (29). Note that factors in month t are portfolios returns with individual stock weights equal to

$$(\hat{\Gamma}' X_{t-1}' W_{t-1} X_{t-1} \hat{\Gamma})^{-1} \hat{\Gamma}' X_{t-1}' W_{t-1},$$

which only depend on information up to time t-1. We adjust factor portfolios by making the portfolios zero-investment through investing in risk-free assets and scaling the long position to be \$1.

In the next subsection, we compare performance of IPCA model and IPCA+DC model with the OS estimates.

5.2. Performance of factors in IPCA and IPCA+DC models

In this subsection, we present IS and OS performance of IPCA factors and IPCA+DC factors.

Table 8 presents results. We order IPCA+DC factors according to the order of clusters presented in the table 1. We order IPCA factors according to standard deviation. Note that in IPCA+DC model when $J = 10$, the corresponding number of clusters is $K = 9$ because IPCA+DC model has an additional market factor, which is the tenth factor (DC10) in table 8. Table shows that IPCA+DC factors outperform IPCA factors. The highest three values of OS Sharpe ratio for IPCA factors are 0.60, 0.59, and 0.53, while those for IPCA+DC factors are 0.81, 0.57 and 0.55, *resp.*

Table 9 presents the correlations between IPCA factors and IPCA+DC factors. It is

shown that factors in each model are not highly correlated, which implies that there is little redundant information in factors. Within IPCA+DC model, correlations between most factors is below 0.4. Moreover, the correlation between IPCA+DC10 factor (the market factor) and IPCA1 factor is 0.99, implying that the market factor plays a role of the first principle component of IPCA. Besides, correlations between most IPCA factors and IPCA+DC factors are not high. Most are below 0.3 (except correlations with IPCA1 factor).

To assess the overall model performance, we compare IPCA and IPCA+DC pair of models with the number of clusters K from 1 to 14. Table 10 shows performance of maximal Sharpe ratio portfolios. In most cases, IPCA+DC models perform better than IPCA models. The average monthly returns of IPCA+DC models are higher when $K = 3, 4, 5, 6, 8, 12, 13, 14$, which range from 0.53% to 0.82%. As for standard deviation, IPCA+DC model has lower values when $K = 3, 4, 5, 6, 7, 8, 9, 11, 12, 13, 14$. The Sharpe ratio deliver perhaps the sharpest differences between IPCA+DC and IPCA models. The Sharpe ratio for IPCA+DC models are higher in all cases except when $K = 1, 2, 10$. The maximal drawdown (MDD) are smaller when $K = 3, 4, 5, 6, 7$. It implies that by adding some clustering structure when estimating factor models, we can get a better spanned pricing model.

Table 11 presents the results of alpha tests. We use the same measure as table 5. IPCA+DC models perform better than IPCA models when K is large. When $K = 9, 11, 12, 13, 14$, the average absolute alpha for IPCA+DC models range from 4.82% to 6.33%, smaller than values achieved by the IPCA models, which range from 6.61% to 7.27%. For more than half of the 94 anomalies, the IPCA+DC models achieve the smallest absolute alpha, compared to less than half of anamalies for IPCA models. Among 94 anomalies, 34, 27, 28, 27, 22 anomalies have alpha with p-value smaller than 0.1 for the IPCA+DC models, compared to 39, 39, 38, 33 and 34 anomalies for the IPCA models. The number of anomalies for which adjusted p-value is smaller than 0.1 and 0.05 follows a similar pattern.

Table 12 uses BS-CZZ method to compare IPCA models and IPCA+DC models with the

same number of clusters k. Table presents posterior probability that each IPCA+DC model holds versus prior expectation κ, which is defined in equation (24). In all cases except the case when $K \leq 2$ or $K = 10$, the posterior probability that IPCA+DC model holds is larger than 0.5. In other words, IPCA+DC model has higher probability to be the right model than IPCA model for most J.

5.3. Cross-sectional return prediction through neural networks with clustering

In this section, we combine clustering with neural networks to predict cross-sectional stock returns. We focus on neural networks for two reasons. First, it is a widely used machine learning method and is found useful in predicting stock returns. Second, different from IPCA, it can capture nonlinear relationship between firm characteristics and stock returns. Previous section shows that clustering can enhance performance in a linear situation, that is IPCA. Consequently, we want to examine whether clustering can enhance accuracy of prediction in a non-linear situation, that is neural network.

5.3.1. Neural networks with clustering

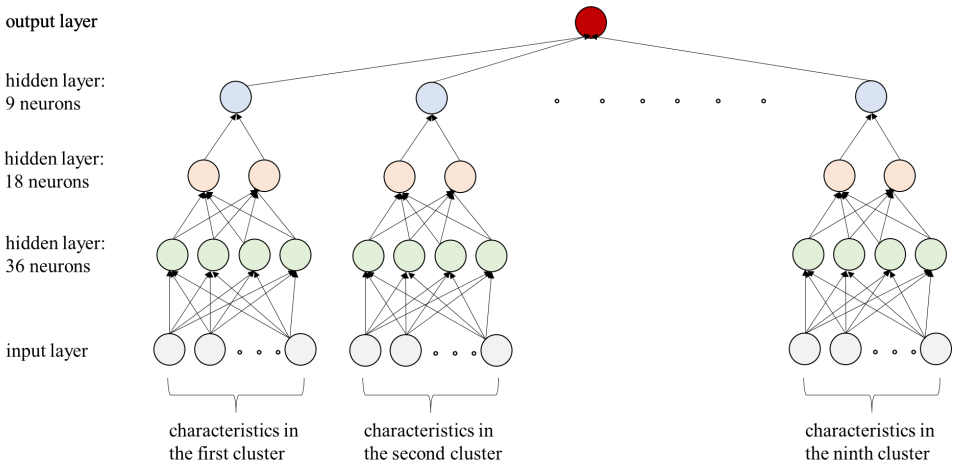
A neural network recognizes the relationship between firm characteristics and stock returns by mimicing the operations of a brain. It contains layers of interconnected neurons (details are in Gu et al. (2020)). Each neuron accepts information through its connection with other neurons and produces a signal by a multiple linear function. Then it feeds the signal into an activation function that may be nonlinear and sends it to other neurons through connection.

We combine clustering with neural networks by limiting connection among neurons. Specifically, we require that only neurons contain information of firm characteristics in the same cluster are connected. Subfigure (a) of figure 3 shows an illustrative example of a neu-

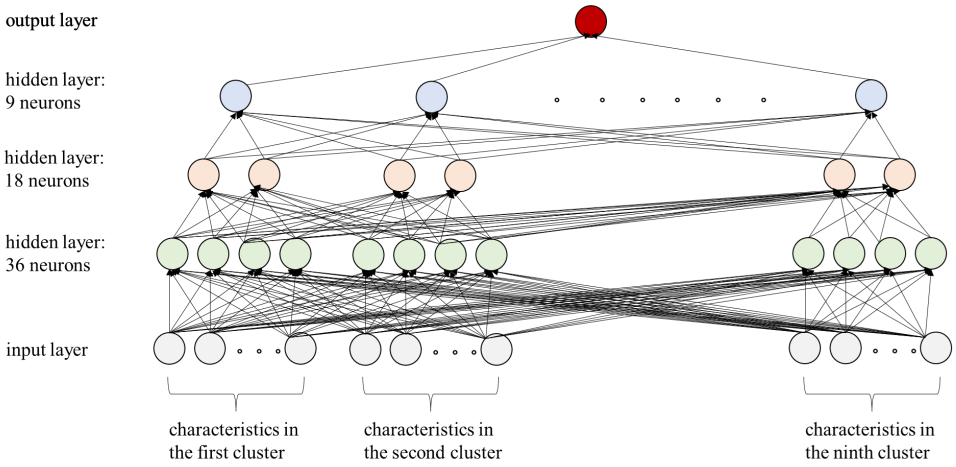
ral network with clustering, which contains three hidden layers with 36, 18 and 9 neurons. Otherwise, a neuron in a network without clustering are connected to all neurons, which is shown in subfigure (b). [2].

As it is impossible to find the optimal network by searching over uncountably many architectures, we fix network architectures ex ante and estimate each of these. We consider three network architectures with 1 hidden layer, which contains 9 neurons. As for activation function, we use rectified linear unit (ReLU) according to Gu et al. (2020). In the following two subsections, we compare performance of neural networks with and without clustering by comparing out-of-sample performance of long-short portfolios and R square.

[2]Similar to Gu et al (2020), our analysis focuses on traditional "feed-forward" neural networks, where information is conveyed in one direction. In a "feed-forward" neural network, a neuron acccepts information only from neurons in the lower layer and sends information only to neurons to the higher layer.



(a) Networks with clustering



(b) Networks without clustering

Figure 3: Neural networks with and without clustering
The figure provides a diagram of neural networks with and without clustering. Grey circles denote the input layer (firm characteristics), and dark red circles denote the output layer.

5.3.2. Performance of long-short portfolios

In this section, we compare performance of neural networks with and without clustering by comparing performance of long-short portfolios.

We construct the long-short portfolios by longing stocks with the highest 20% predicted returns and shorting stocks with the lowest 20% predicted returns. To make the performance out-of-sample, we divided the 37 years of data into 18 years of training sample (Jan.

1985 - Dec. 2002) and the reserved sample, and examine performance in the reserved sample. Performance varies with hyperparameters as they control model complexity (Gu et al., 2020). To show robustness of the results, performances under various hyperparameters are presented.

Panel A of table 13 shows results. Each row shows performance under a hyperparameter value. Hyper-parameters include learning rate (lr), patience of early stopping algorithm (p), and batch size ($batch\_size$). Illustration of those hyper-parameters can be found in Gu et al. (2020). The learning rate is fixed at 0.001, and other hyperparameters are in a grid of values $batch_size = \{100000, 15000\}, p = \{10, 20, 30\}$. Panel A of table 13 shows that the networks with clustering have higher Sharpe ratio than the networks without clustering in most cases. For example, when batch size is 10000 and patience is 20, the Sharpe ratio for network with clustering is 0.56, while is 0.35 for network without clustering. The pattern is similar when hyperparameters take other values. If we choose the optimal hyperparameter with R square in the valiadation sample, the optimal hyperparameter is $\{batch_size = 15000, patience = 10\}$ for both networks with and without clustering. When the networks have 1 hidden layer, the networks with clustering achieve higher Sharpe ratios when patience is larger than 10. In those cases, the Sharpe ratios for networks with clustering are 1.70, 1.90, 1.75 and 1.79 for different hyperparameters. While the networks with clustering achieve 1.16, 1.42, 1.72 and 1.64. The gap widen when the networks have 2 or 3 hidden layers. When the networks have 2 or 3 hidden layers, the network with clustering have higher Sharpe ratios under all values of hyperparameters.

5.3.3. R-squared

In this section, we compare performance of neural networks with and without clustering by comparing R-squared.

According to Gu et al. (2020), R-squared serves as the objective function when training

models.

$$R^2 = 1 - \frac{\sum_{t=1}^{T}\sum_{i=1}^{N}(r_{n,t} - \hat{r}_{n,t})^2}{\sum_{t=1}^{T}\sum_{i=1}^{N} r_{n,t}^2}, \tag{31}$$

where $r_{n,t}$ is stock return of firm n in month t and $\hat{r}_{n,t}$ is predicted stock return.

Table 14 shows results. The networks with clustering have higher R square than the networks without clustering. Specifically, the networks with clustering achieve positive R square in most cases while the networks without clustering achieve negative R square.

6. Robustness Tests

In previous results, we use full sample and do not drop small firms. To avoid that the results are driven by small firms, we construct portfolios in equation (11) with NYSE breakpoints and repeat all tests. Results are in table 15 to 17 and all results are robust.

Table 15 presents performance of factors constructed with NYSE breakpoints. Results are consistent with the result we get when construct factors with full-sample breakpoints. By splitting $Trading\ frictions$ and $Momentum$ in IC, we get factors with higher Sharpe ratios. Table 16 shows performance of OS maximal Sharpe ratio portfolios. DC-Model has the highest average monthly return, the lowest standard deviation and thus the highest Sharpe ratio. In addition, it has the lowest MDD. Table 17 applies the BS-CZZ method and reports posterior probability $Pr(M_i|D)$ versus prior expectation κ in equation (24). It is shown that the DC-Model has the highest probability to be the true model whether compared with each of benchmarks or all benchmarks simultaneously.

7. Conclusion

In this paper, we propose a new approach to construct a factor model from firm characteristics. Similar to IPCA model of Kelly et al. (2019), our model allows for latent factors

and time-varying loadings by introducing observable characteristics that instrument for the unobservable dynamic loadings. Instead of using principal component analysis to reduce the dimensionality of the return covariance matrix, we use a version of hierarchical cluster analysis to blocklize the cross-sectional firm characteristics covariance matrix, and hence effectively reduce the dimension of the factor model. As each of our factors are correlated to a cluster of firm characteristics, our factors are economically interpretable. Besides, we find the optimal number of clusters endogeneously.

The factor model based on our approach outperforms other benchmark models through various statistical and model comparison tests. Furthermore, viewed as a preprocessor of data, our approach complements IPCA and machine learning methods.

References

Barra. 1998. Global Equity - Risk Model. Handbook.

Barillas, F., and Shanken, J. 2018. Comparing asset pricing models. *Journal of Finance* 73:715-754.

Bojchevski, A., Matkovic, Y., and Günnemann, S. 2017. Robust spectral clustering for noisy data: Modeling sparse corruptions improves latent embeddings. In Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining.

Carhart, M. 1997. On persistence in mutual fund performance. *Journal of Finance* 52:57-82.

Chib, S., Zeng, X., and Zhao, L. 2020. On comparing asset pricing models. *Journal of Finance* 75:551-577.

Cochrane, J. H. 2011. Presidential address: Discount rates. *The Journal of finance* 66:1047-1108.

Fama, E., and French, K. 1993. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* 33:3-56.

Fama, E., and French, K. 2015. A five-factor asset pricing model. *Journal of Financial Economics* 116:1-22.

Fama, E. F., and MacBeth, J. D. 1973. Risk, return, and equilibrium: Empirical tests. *Journal of political economy* 81:607-636.

Freyberger, J., Neuhierl, A., and Weber, M. 2020. Dissecting characteristics nonparametrically. *Reivew of Financial Studies* 33:2326–2377.

Gu, S., Kelly, B., and Xiu, D. 2020. Empirical asset pricing with machine learning. *Review of Financial Studies* 33:2223-2273.

Han, Y., He, A., Rapach, D. and Zhou, G. 2021. Firm characteristics and expected stock returns. Available at SSRN.

Harvey, C. R., Liu, Y., and Zhu, H. 2016. ... and the cross-section of expected returns. *Review of Financial Studies* 29:5–68.

Hansen, Lars Peter, and Ravi Jagannathan. 1991. Implications of Security Market Data for Models of Dynamic Economies. *Journal of Political Economy* 99:225–262.

Hastie, T., Tibshirani, R., and Friedman, J. 2009. The elements of statistical learning. 2nd Ed.

Hou, K., Xue, C., and Zhang, L. 2015. Digesting anomalies: An investment approach. *Review of Financial Studies* 28:650-705.

Karypis, G., Han, E., and Kumar, V. 1999. CHAMELEON a hierarchical clustering algorithm using dynamic modeling. *Computer* 32:68-75.

Kelly B, Pruitt S, and Su Y. 2019. Characteristics are covariances: a unified model of risk and return. *Journal of Financial Economics* 134:501–24.

Kozak, S., Nagel, S., and Santosh, S. 2020. Shrinking the cross-section. *Journal of Financial Economics* 135:271-292.

Lee, H., and Potscher, M. 2005 Model selection and inference: facts and friction. *Econometric Theory* 21:21-59.

Lewellen, J. 2015. The cross-section of expected stock returns. *Critical Finance Review* 4:1-44.

Stambaugh, R., and Yuan, Y. 2017. Mispricing factors. *The Review of Financial Studies* 30:1270-1315.

A. Appendix

In this appendix, we prove the convergence of our method. Denote $\hat{\beta}_{kt} = (\hat{\beta}_{0,k,t}, \hat{\beta}_{1,k,t}, ..., \beta_{I_k,k,t})'$ with I_k the number of characteristics in the k-th cluster. As Equation (5), for each cluster k, normalized $\hat{X}_t = (\hat{X}_{1t}, ..., \hat{X}_{I_k,t})$,

$$\hat{X}_{it} = F_{kt} + e_{i,k,t} \tag{A1}$$

with $e_{i,k,t}$ the homogeneous measurement errors with standard deviation $\sigma^e_{k,t}$. Here we assume the cross-sectional standard deviation of measurement errors are the same across the characteristics $i = 1, ..., I_k$ within each cluster. Let σ_{F_k} be the cross-sectional standard deviation for the single hidden factor $F_{k,t}$. The first step OLS regression, Equation (9), can be written as

$$\begin{aligned} R^n_t &= \hat{\beta}_{0,k,t} + \sum_{i=1}^{I_k} \hat{\beta}_{i,k,t}(F^n_{k,t-1} + e^n_{i,k,t-1}) + \epsilon^n_t \\ &= \hat{\beta}_{0,k,t} + \sum_{i}^{I_k} \hat{\beta}_{i,k,t} F^n_{k,t-1} + \sum_{i=1}^{I_k} \hat{\beta}_{i,k,t} e^n_{i,k,t-1} + \epsilon^n_t. \end{aligned} \tag{A2}$$

The OLS predictive regression consists of measurement error term $\sum_{i=1}^{I_k} \hat{\beta}_{i,k,t} e^n_{i,k,t-1}$. We will show that this variance of this measurement error term decreases with I_k under certain technical conditions.

Assume the unbiased estimator $\hat{\beta}^0_{k,t}$ for

$$R^n_t = \hat{\beta}^0_{0,k,t} + \hat{\beta}^0_{k,t} F^n_{k,t-1} + \nu^n_t. \tag{A3}$$

Obviously,

$$|\sum_{i=1}^{I_k} \hat{\beta}_{i,k,t}| < |\hat{\beta}^0_{k,t}|.$$

We assume the cross-sectional standard deviation of F_k is larger than that of measurement

errors, σ_k^e, which guarantees that the factor has the largest eigenvalue and $\hat{\beta}_{i,k,t}$ are of the same sign for all $i = 1, ..., I_k$. Hence

$$\sum_{i=1}^{I_k} |\hat{\beta}_{i,k,t}| < |\hat{\beta}_{k,t}^0|,$$

and the cross-sectional variance of measurement error in Equation (A2) is given by

$$Var(\sum_{i=1}^{I_k} \hat{\beta}_{i,k,t} e_i) = \sum_{i=1}^{I_k} \hat{\beta}_{i,k,t}^2 \sigma_e^2 \leq (\sum_{i=1}^{I_k} \hat{\beta}_{i,k,t})^2 \sigma_e^2 \leq (\hat{\beta}_{k,t}^0)^2 \sigma_e^2, \tag{A4}$$

which decreases monotonically with I_k. This says that when there is measurement error in explanatory variable, the more independent measurements the better. The additional explanatory variables serve as instrumental variables. Note For $I_k \to +\infty$, the measurement error variance disappears. Hence, our 3-step regression approach is consistent.

Table 1: DC and IC

Table presents DC and IC. Panel A presents number of firm characteristics in each cluster of DC and IC. The darkness of color represents how large the value is. For example, value in the first row and first column is 5, meaning that there are 5 firm characteristics in IC1 cluster and in DC1 cluster. We use DC1-DC9 as identification (ID) for 9 clusters in DC, and IC1-IC6 as ID for 6 clusters in IC. Clusters corresponding to those IDs are in Panel B and C.

Panel A: # firm characteristics in DC and IC

	DC1	DC2	DC3	DC4	DC5	DC6	DC7	DC8	DC9
IC1	5	5	4	2	0	0	0	0	0
IC2	0	0	0	0	3	2	0	4	0
IC3	0	0	0	0	0	0	8	0	0
IC4	0	0	0	0	0	0	9	4	0
IC5	0	0	0	0	0	0	18	15	3
IC6	0	0	0	0	0	0	3	9	0

Panel B: clusters in DC

Cluster ID	Cluster
DC1	Trading frictions (measured by volume)
DC2	Illiquidity
DC3	Trading frictions (measured by return)
DC4	Beta
DC5	Long-run momentum
DC6	Short-run momentum
DC7	Profitability
DC8	Growth
DC9	Accruals

Panel C: clusters in IC

Cluster ID	Cluster
IC1	Trading frictions
IC2	Momentum
IC3	Value
IC4	Profitability
IC5	Intangible
IC6	Investment

Table 2: Performance of DC and IC factors

Table presents performance of factors in DC-Model and IC-Model. Panel A reports performance of factors constructed on firm characteristics in *Trading frictions* in IC. Those factors include trading frictions factor in IC-Model and 4 factors in DC-Model. Panel B reports performance of factors constructed on firm characteristics in *Momentum* in IC. Those factors include momentum factor in IC-Model, long-run momentum factor and short-run momentum factor in DC-Model. Panel C reports performance of other factors in IC-Model and DC-Model. Performance measures include sample mean, sample standard deviation, annualized Sharpe ratio, and maximum drawdown (MDD) of monthly return. The sample period is from Jan. 1990 through Dec. 2021.

Cluster		Mean(%)	S.D.(%)	Sharpe	MDD(%)
Panel A: Trading frictions cluster					
IC1	*Trading frictions*	0.03	1.55	0.07	24.50
DC1	*Trading frictions (measured by volume)*	0.10	1.05	0.32	15.17
DC2	*Illiquidity*	0.19	1.94	0.34	22.96
DC3	*Trading frictions (measured by return)*	0.10	3.00	0.11	51.21
DC4	*Beta*	0.11	2.28	0.16	25.84
Panel B: Momentum cluster					
IC2	*Momentum*	0.26	2.10	0.42	28.40
DC5	*Short − run momentum*	0.17	1.58	0.36	20.49
DC6	*Long − run momentum*	0.38	1.57	0.83	13.66
Panel C: Other clusters					
IC3	*Value*	0.19	1.54	0.43	22.35
IC4	*Profitability*	0.22	1.11	0.70	12.87
IC5	*Intangibles*	0.14	1.24	0.40	11.07
IC6	*Investment*	0.10	1.13	0.32	18.65
DC7	*Profitability*	0.27	1.12	0.85	10.38
DC8	*Growth*	0.09	1.05	0.29	23.38
DC9	*Accruals*	0.08	1.32	0.20	28.89

Table 3: Correlation of factors in DC-Model and IC-Model

Table presents correlations of factors in DC-Model and IC-Model. Each column or each row represents a factor corresponding to a cluster among DC1-DC9 and IC1-IC6. Clusters corresponding to DC1-DC9 and IC1-IC6 are in the bottom of table 1. The sample period is from Jan. 1990 through Dec. 2021.

		factors in C model									factors in NC model					
		1	2	3	4	5	6	7	8	9	1	2	3	4	5	6
	1	1.00	0.16	0.19	-0.06	0.02	0.06	0.04	0.19	-0.12	0.04	0.04	0.07	-0.02	0.13	0.19
	2	0.16	1.00	0.44	0.16	-0.05	0.23	0.09	0.15	-0.20	-0.54	0.06	0.27	0.28	0.19	-0.10
	3	0.19	0.44	1.00	0.20	0.01	0.21	-0.10	0.27	-0.43	-0.42	-0.09	0.59	0.32	0.30	-0.25
	4	-0.06	0.16	0.20	1.00	-0.08	-0.18	-0.02	0.03	-0.15	-0.40	-0.22	0.13	0.19	0.02	-0.11
factors in	5	0.02	-0.05	0.01	-0.08	1.00	-0.01	-0.21	0.03	0.03	0.13	0.75	0.02	-0.10	-0.09	-0.05
C model	6	0.06	0.23	0.21	-0.18	-0.01	1.00	0.07	-0.04	0.08	-0.11	0.40	0.14	0.09	0.09	-0.10
	7	0.04	0.09	-0.10	-0.02	-0.21	0.07	1.00	0.08	-0.12	0.05	-0.03	-0.12	0.44	0.35	0.22
	8	0.19	0.15	0.27	0.03	0.03	-0.04	0.08	1.00	-0.17	-0.08	0.08	-0.01	0.13	0.54	0.31
	9	-0.12	-0.20	-0.43	-0.15	0.03	0.08	-0.12	-0.17	1.00	0.01	0.16	-0.34	-0.40	0.02	0.01
	1	0.04	-0.54	-0.42	-0.40	0.13	-0.11	0.05	-0.08	0.01	1.00	0.05	-0.19	-0.11	-0.18	0.16
	2	0.04	0.06	-0.09	-0.22	0.75	0.40	-0.03	0.08	0.16	0.05	1.00	-0.06	-0.15	0.01	0.08
factors in	3	0.07	0.27	0.59	0.13	0.02	0.14	-0.12	-0.01	-0.34	-0.19	-0.06	1.00	0.26	0.10	-0.26
NC model	4	-0.02	0.28	0.32	0.19	-0.10	0.09	0.44	0.13	-0.40	-0.11	-0.15	0.26	1.00	0.06	-0.10
	5	0.13	0.19	0.30	0.02	-0.09	0.09	0.35	0.54	0.02	-0.18	0.01	0.10	0.06	1.00	0.04
	6	0.19	-0.10	-0.25	-0.11	-0.05	-0.10	0.22	0.31	0.01	0.16	0.08	-0.26	-0.10	0.04	1.00

Table 4: Performance of maximal Sharpe ratio portfolios

The table reports performance of maximal Sharpe ratio portfolios for DC-Model and factor models, including IC-Model, FF3, Car4, Q4, FF5, and cross-sectional stock return prediction method, including Lasso, Ridge and Enet. For 6 factor models, The returns of maximal Sharpe ratio portfolios are constructed on a purely out-of-sample basis by using the mean and covariance matrix of estimated factors through t and tracking the post-formation t + 1 return. For cross-sectional return prediction method, the returns of portfolios are constructed based on return prediction with panel regression using data through t, where hyper-parameters are chosen with cross-validation, and tracking the post-formation t + 1 return. In the portfolios, 50% stocks with the highest return prediction are in the long position and the rest are in the short position. Performance measures include sample mean, sample standard deviation, annualized Sharpe ratio and maximum drawdown (MDD) of monthly return. The sample period is from Jan. 1990 through Dec. 2021.

Models	Performance of OS maximal Sharpe ratio portfolio			
	Mean(%)	S.D.(%)	Sharpe	MDD(%)
DC	0.53	1.38	1.34	8.22
IC	0.37	1.43	0.90	17.88
FF3	0.37	2.60	0.49	33.38
Car4	0.42	2.67	0.55	30.39
Q4	0.31	1.60	0.68	22.68
FF5	0.40	1.51	0.91	17.12
Lasso	0.33	4.22	0.27	36.45
Ridge	0.47	4.71	0.35	40.29
Enet	0.41	4.21	0.34	36.41

Table 5: Alpha Test

The table reports results of alpha tests. For model l, We run regression $f_t = \alpha_l + \beta_l F_{l,t} + e_{l,t}$, where $f_t \in R^N$ are test portfolios returns in month t, $F_{l,t}$ are factors in model l in month t. We test 6 models, DC-Model, IC-Model, FF3, Car4, Q4 and FF5. We use 2 sets of f_t in Panel A and Panel B, respectively. In Panel A, the test portfolios returns are 94 anomalies. In Panel B, the test portfolio returns are 47 anomalies: For each factor in DC and IC models, the five anomalies which are most highly correlated with the factor are eliminated. This procedure leaves 47 anomalies in Panel B. The table reports the average absolute alpha, the number of anomalies for which DC model produces the smallest absolute alpha among the models being compared in the table, the number of anomalies for which p-value is smaller than 0.1, and 0.05, the number of anomalies for which adjusted p-value is smaller than 0.1, and 0.05. Adjusted p-value is adjusted for 5-lags auto-correlation in error term $\{e_{l,t}\}_{t=1}^{T}$. The sample period is from Jan. 1990 through Dec. 2021.

	DC	IC	FF3	Car4	Q4	FF5
Panel A. 94 anomalies						
Avg. $\|\alpha\|$ (%)	0.08	0.10	0.14	0.15	0.13	0.13
# min $\|\alpha\|$	27	22	16	10	9	10
# $p < 0.1$	17	22	31	39	38	52
# $p < 0.05$	5	13	21	32	28	39
# $p_{adj} < 0.1$	11	17	26	34	38	47
# $p_{adj} < 0.05$	6	8	18	22	26	36
Panel B. 47 anomalies						
Avg. $\|\alpha\|$ (%)	0.22	0.24	0.26	0.26	0.25	0.24
# min $\|\alpha\|$	15	7	2	10	4	9
# $p < 0.1$	19	22	21	23	26	25
# $p < 0.05$	16	17	20	19	24	22
# $p_{adj} < 0.1$	16	16	21	21	23	23
# $p_{adj} < 0.05$	15	13	18	19	21	18

Table 6: BS-CZZ method: posterior probability versus prior expectation

The table applies the BS-CZZ method and compares DC-Model with benchmarks. It reports posterior probability $Pr(M_l|D)$ versus prior expectation κ. Panel A compares the DC-Model with each of benchmarks and reports posterior probability that the DC-Model holds. Panel B compares all models simultaneously and reports posterior probability that each model holds. The sample period is from Jan. 1990 through Dec. 2021.

	$\kappa = 2.0$	2.2	2.4	2.6	2.8	3.0	3.2	3.4	3.6	3.8	4.0
Panel A: Compare the DC-Model with each of benchmarks											
IC	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
FF3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Car4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Q4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
FF5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Panel B: Compare the DC-Model with benchmarks simultaneously											
DC	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
IC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FF3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Car4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FF5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 7: OS Clustering: Performance of maximal Sharpe ratio portfolios

The table reports performance of maximal Sharpe ratio portfolios for OSDC-Model and factor models, including IC-Model, FF3, Car4, Q4, FF5, and cross-sectional stock return prediction method, including Lasso, Ridge and Enet. For 6 factor models, The returns of maximal Sharpe ratio portfolios are constructed on a purely out-of-sample basis by using the mean and covariance matrix of estimated factors through t and tracking the post-formation t + 1 return. For cross-sectional return prediction method, the returns of portfolios are constructed based on return prediction with panel regression using data through t, where hyper-parameters are chosen with cross-validation, and tracking the post-formation t + 1 return. In the portfolios, 50% stocks with the highest return prediction are in the long position and the rest are in the short position. Performance measures include sample mean, sample standard deviation, annualized Sharpe ratio and maximum drawdown (MDD) of monthly return. The sample period is from Jan. 2012 through Dec. 2021.

Model	Performance of maximal Sharpe ratio portfolio			
	Mean(%)	S.D.(%)	Sharpe	MDD(%)
OSDC	0.51	1.36	1.30	9.29
IC	0.30	1.47	0.71	10.28
FF3	0.62	2.60	0.83	21.50
Q4	0.63	2.25	0.97	32.25
Car4	0.43	1.92	0.77	22.09
FF5	0.42	1.58	0.92	21.06
Lasso	0.05	2.61	0.07	16.38
Ridge	0.15	2.84	0.18	14.79
Enet	0.12	2.58	0.16	12.68

Table 8: Performance of factors in IPCA and IPCA+DC

Table presents performance of factors in IPCA and IPCA+DC models. Each row for IPCA+DC shows performance of factors corresponding to the cluster ID. The relationship between cluster and cluster ID is in Panel B of table 1. IPCA+DC10 represent the factor to which exposure is not correlated to firm characteristics. IPCA1 - IPCA10 represent 10 factors in IPCA model and are ordered according to standard deviation. Performance measures include sample mean, sample standard deviation, annualized Sharpe ratio, and maximum drawdown (MDD) of monthly return. The sample period is from Jan. 1990 through Dec. 2021.

	IPCA+DC				IPCA			
Factor	Mean(%)	S.D.(%)	Sharpe	MDD(%)	Mean(%)	S.D.(%)	Sharpe	MDD(%)
1	0.06	1.93	0.12	37.26	0.72	4.92	0.51	58.83
2	0.09	1.99	0.16	45.78	0.34	2.21	0.53	23.33
3	0.47	2.99	0.55	35.70	0.21	1.90	0.39	26.99
4	0.13	2.97	0.16	37.92	0.11	1.94	0.20	38.71
5	0.27	2.57	0.37	24.92	0.25	2.15	0.40	38.77
6	0.74	3.19	0.81	40.23	0.11	1.79	0.21	22.27
7	0.19	2.27	0.28	43.25	0.22	1.61	0.47	15.42
8	0.13	1.71	0.27	19.74	0.37	2.15	0.59	32.52
9	0.05	1.29	0.14	19.44	0.34	1.98	0.60	17.53
10	0.74	4.48	0.57	53.70	0.01	1.97	0.02	43.16

Table 9: Correlation of factors in IPCA and IPCA+DC

Table presents correlations of factors in IPCA and IPCA+DC models with OS estimates. Each column or each row represents a factor corresponding to the cluster ID. The relationship between cluster and cluster ID is in Panel B of table 1. IPCA+DC10 represent the factor to which exposure is not correlated to firm characteristics. IPCA1 - 10 represent 10 factors in IPCA model. The sample period is from Jan. 1990 through Dec. 2021.

	IPCA+DC										IPCA									
	IPCA+DC1	2	3	4	5	6	7	8	9	10	IPCA1	2	3	4	5	6	7	8	9	10
IPCA+DC1	1.00	0.12	-0.41	0.17	0.14	-0.01	0.13	0.07	0.06	0.33	0.33	-0.14	0.06	-0.24	0.00	0.07	0.03	-0.28	-0.02	-0.21
2	0.12	1.00	-0.05	0.05	-0.01	0.08	0.14	0.27	-0.02	0.05	0.05	-0.11	0.02	0.00	-0.03	0.11	0.17	-0.22	0.16	0.12
3	-0.41	-0.05	1.00	-0.26	-0.19	0.01	0.07	0.14	-0.06	-0.51	-0.52	0.23	0.07	0.09	0.19	0.16	-0.07	0.35	0.10	0.39
4	0.17	0.05	-0.26	1.00	0.23	-0.17	-0.09	-0.04	0.16	0.57	0.63	-0.04	0.04	-0.15	0.03	-0.23	-0.40	-0.12	0.06	-0.07
5	0.14	-0.01	-0.19	0.23	1.00	-0.61	-0.18	0.01	-0.01	0.31	0.32	-0.25	-0.02	0.04	0.09	-0.05	-0.09	0.48	-0.17	0.11
6	-0.01	0.08	0.01	-0.17	-0.61	1.00	0.15	0.10	-0.02	-0.11	-0.14	0.36	0.18	0.15	0.01	0.17	0.31	-0.39	0.25	-0.34
7	0.13	0.14	0.07	-0.09	-0.18	0.15	1.00	0.12	0.01	-0.13	-0.14	-0.01	0.23	-0.06	-0.35	0.26	0.17	-0.22	0.37	-0.01
8	0.07	0.27	0.14	-0.04	0.01	0.10	0.12	1.00	-0.05	0.00	-0.03	-0.05	0.07	0.08	-0.02	0.07	0.27	0.01	0.04	-0.05
9	0.06	-0.02	-0.06	0.16	-0.01	-0.02	0.01	-0.05	1.00	0.01	0.01	-0.14	0.13	-0.18	0.02	-0.14	-0.05	-0.06	0.09	-0.01
10	0.33	0.05	-0.51	0.57	0.31	-0.11	-0.13	0.00	0.01	1.00	0.99	-0.14	-0.02	-0.12	0.01	-0.14	0.06	-0.03	-0.12	-0.11
IPCA1	0.33	0.05	-0.52	0.63	0.32	-0.14	-0.14	-0.03	0.01	0.99	1.00	-0.13	-0.03	-0.13	0.03	-0.16	-0.05	-0.03	-0.11	-0.11
2	-0.14	-0.11	0.23	-0.04	-0.25	0.36	-0.01	-0.05	-0.14	-0.14	-0.13	1.00	0.15	-0.01	0.00	-0.14	-0.05	0.03	0.02	-0.08
3	0.06	0.02	0.07	0.04	-0.02	0.18	0.23	0.07	0.13	-0.02	-0.03	0.15	1.00	-0.14	-0.13	-0.02	0.03	0.00	0.01	-0.12
4	-0.24	0.00	0.09	-0.15	0.04	0.15	-0.06	0.08	-0.18	-0.12	-0.13	-0.01	-0.14	1.00	-0.10	0.09	0.13	0.02	-0.02	0.03
5	0.00	-0.03	0.19	0.03	0.09	0.01	-0.35	-0.02	0.02	0.01	0.03	0.00	-0.13	-0.10	1.00	-0.03	-0.16	0.17	-0.09	0.08
6	0.07	0.11	0.16	-0.23	-0.05	0.17	0.26	0.07	-0.14	-0.14	-0.16	-0.14	-0.02	0.09	-0.03	1.00	0.19	0.01	0.17	-0.03
7	0.03	0.17	-0.07	-0.40	-0.09	0.31	0.17	0.27	-0.05	0.06	-0.05	-0.05	0.03	0.13	-0.16	0.19	1.00	-0.15	0.04	-0.10
8	-0.28	-0.22	0.35	-0.12	0.48	-0.39	-0.22	0.01	-0.06	-0.03	-0.03	0.03	0.00	0.02	0.17	0.01	-0.15	1.00	-0.21	0.23
9	-0.02	0.16	0.10	0.06	-0.17	0.25	0.37	0.04	0.09	-0.12	-0.11	0.02	0.01	-0.02	-0.09	0.17	0.04	-0.21	1.00	0.00
10	-0.21	0.12	0.39	-0.07	0.11	-0.34	-0.01	-0.05	-0.01	-0.11	-0.11	-0.08	-0.12	0.03	0.08	-0.03	-0.10	0.23	0.00	1.00

Table 10: Performance of maximal Sharpe ratio portfolio of IPCA+DC and IPCA models

The table compares maximal Sharpe ratio portfolios performance of IPCA+DC model and IPCA model for each number of cluster K. Performance measures include sample mean, sample standard deviation, annualized Sharpe ratio and maximum drawdown (MDD) of monthly return. The sample period is from Jan. 1990 through Dec. 2021.

K	IPCA+DC				IPCA			
	Mean (%)	S.D. (%)	Sharpe	MDD (%)	Mean (%)	S.D. (%)	Sharpe	MDD (%)
1	0.49	4.46	0.38	61.77	0.58	2.90	0.69	40.97
2	0.65	3.50	0.64	43.84	0.73	2.90	0.87	37.52
3	0.74	2.37	1.09	30.38	0.58	2.83	0.71	33.29
4	0.70	2.18	1.11	21.97	0.54	2.60	0.72	37.08
5	0.72	2.09	1.19	17.74	0.64	2.37	0.94	25.71
6	0.65	2.00	1.12	18.31	0.61	2.29	0.93	25.29
7	0.56	1.75	1.11	20.31	0.57	2.08	0.94	23.53
8	0.53	1.76	1.05	18.88	0.53	2.05	0.89	18.30
9	0.71	1.68	1.46	22.25	0.72	1.86	1.34	22.08
10	0.73	1.68	1.50	19.91	0.75	1.59	1.64	13.39
11	0.71	1.44	1.72	16.68	0.72	1.55	1.61	11.65
12	0.78	1.51	1.79	17.45	0.71	1.54	1.61	11.85
13	0.82	1.62	1.75	17.54	0.73	1.49	1.69	12.01
14	0.78	1.46	1.85	15.42	0.67	1.50	1.54	12.61

Table 11: Alpha test between IPCA+DC and IPCA models

The table uses alpha test to compare IPCA and IPCA+DC model with each number of clusters K. We run regression $f_t = \alpha_l + \beta_l F_{l,t} + e_{l,t}$, where f_t are 94 managed portfolio returns and F_t are model factors. For each model, the table reports the average absolute alpha, the number of anomalies for which IPCA+DC produces the smaller absolute alpha, the number of anomalies for which p-value is smaller than 0.1, and 0.05, the number of anomalies for which adjusted p-value is smaller than 0.1, and 0.05. Adjusted p-value is adjusted for 5-lags auto-correlation in error term $\{e_{l,t}\}_{t=1}^{T}$. The sample period is from Jan. 1990 through Dec. 2021.

Statistics	Model	K													
		1	2	3	4	5	6	7	8	9	10	11	12	13	14
Average $\|\alpha\|$ (%)	IPCA+DC	10.21	9.83	10.08	10.08	10.76	10.75	9.22	8.32	6.33	6.54	5.29	5.33	5.36	4.82
	IPCA	8.04	6.94	6.97	7.40	6.45	6.45	6.34	6.70	6.61	6.47	7.27	7.03	6.77	6.66
# min $\|\alpha\|$	IPCA+DC	30	23	25	20	21	19	28	31	56	54	56	58	62	57
	IPCA	64	71	69	74	73	75	66	63	38	40	38	36	32	37
# $p < 0.1$	IPCA+DC	57	55	56	57	56	57	52	45	34	36	27	28	27	22
	IPCA	54	35	38	39	36	34	34	38	39	28	39	38	33	34
# $p < 0.05$	IPCA+DC	45	48	50	49	52	52	47	40	26	25	23	17	18	13
	IPCA	43	27	32	34	29	24	23	29	32	20	29	30	24	25
# $p_{adj} < 0.1$	IPCA+DC	53	50	48	46	50	49	48	39	31	31	24	26	24	18
	IPCA	51	34	35	36	35	30	30	34	37	24	34	36	28	28
# $p_{adj} < 0.05$	IPCA+DC	43	37	36	37	37	39	39	34	21	20	22	15	17	12
	IPCA	38	28	29	32	29	24	21	26	24	18	24	27	21	23

Table 12: BS-CZZ method for IPCA: posterior probability versus prior expectation

The table uses BS-CZZ method to compare IPCA and IPCA+DC model for each number of clusters K. Table presents posterior probability that IPCA+DC model holds versus prior expectation κ, which is defined in equation (24). The sample period is from Jan. 1990 through Dec. 2021.

K	κ										
	2	2.2	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4
1	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
2	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03
3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
6	0.86	0.86	0.87	0.87	0.87	0.87	0.87	0.88	0.88	0.88	0.88
7	0.99	0.99	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00
8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
9	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
10	0.07	0.06	0.06	0.06	0.06	0.05	0.05	0.05	0.05	0.05	0.05
11	0.98	0.98	0.98	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
12	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
13	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
14	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 13: Sharpe of long-short portfolios for neural networks with and without clustering

The table reports Sharpe ratio of long-short portfolios for neural networks with and without clustering. We compare 6 models in total: networks with or without clustering and 1-3 hidden layers. Each row shows performance under a hyperparameter value. Hyper-parameters include learning rate (lr), patience of early stopping algorithm (p), and batch size ($batch\_size$). Illustration of those hyper-parameters can be found in Gu et al. (2020). The learning rate is fixed at 0.001, and other hyperparameters are in a grid of values $batch_size = \{100000, 15000\}, p = \{10, 20, 30\}$. The sample period is from Jan. 2012 through Dec. 2021.

Hyper-parameter		Without clustering			With clustering		
batch size	patience	NN3	NN4	NN5	NN3	NN4	NN5
10000	10	1.35	1.46	1.38	1.07	1.80	1.83
10000	20	1.16	1.66	1.27	1.70	2.21	1.96
10000	30	1.42	1.43	1.32	1.90	2.31	2.26
15000	10	1.75	1.37	1.45	1.38	1.75	1.87
15000	20	1.72	1.38	1.20	1.75	2.15	1.91
15000	30	1.64	1.31	1.79	1.79	2.10	2.16

Table 14: R square for neural networks with and without clustering

The table reports out-of-sample R square for neural networks with and without clustering. We compare 6 models in total: networks with or without clustering and 1-3 hidden layers. Each row shows performance under a hyperparameter value. Hyper-parameters include learning rate (lr), patience of early stopping algorithm (p), and batch size ($batch\_size$). Illustration of those hyper-parameters can be found in Gu et al. (2020). The learning rate is fixed at 0.001, and other hyperparameters are in a grid of values $batch_size = \{100000, 15000\}, p = \{10, 20, 30\}$. The sample period is from Jan. 2012 through Dec. 2021.

Hyper-parameter		Without clustering			With clustering		
batch size	patience	NN3	NN4	NN5	NN3	NN4	NN5
10000	10	-0.32	-0.45	-0.28	0.28	0.11	0.27
10000	20	-0.56	-0.52	-0.60	0.16	0.04	-0.01
10000	30	-0.32	-0.68	-0.69	0.08	-0.15	-0.09
15000	10	-0.24	-0.54	-0.30	0.40	0.17	0.19
15000	20	-0.29	-0.79	-0.66	0.31	0.12	0.39
15000	30	-0.38	-1.08	-1.29	0.30	0.11	0.14

Table 15: Performance of DC and IC factors with NYSE breakpoints

Table presents performance of factors in DC-Model and IC-model constructed with NYSE breakpoints. Panel A reports performance of factors constructed on firm characteristics in *Trading frictions* in IC. Those factors include trading frictions factor in IC-Model and 4 factors in DC-Mmodel. Panel B reports performance of factors constructed on firm characteristics in *Momentum* in IC. Those factors include momentum factor in IC-Model, long-run momentum factor and short-run momentum factor in DC-Model. Panel C reports performance of other factors in IC-Model and DC-Model. Performance measures include sample mean, sample standard deviation, annualized Sharpe ratio, and maximum drawdown (MDD) of monthly return. The sample period is from Jan. 1990 through Dec. 2021.

	Cluster	Mean(%)	S.D.(%)	Sharpe	MDD(%)
Panel A: Trading frictions cluster					
IC1	*Trading frictions*	0.00	1.35	0.00	20.08
DC1	*Trading frictions (measured by volume)*	0.05	0.99	0.16	15.59
DC2	*Illiquidity*	0.09	1.66	0.19	22.21
DC3	*Trading frictions (measured by return)*	0.01	1.85	0.02	36.38
DC4	*Beta*	0.14	2.53	0.19	27.50
Panel B: Momentum cluster					
IC2	*Momentum*	0.25	2.01	0.44	31.35
DC5	*Short − run momentum*	0.38	1.57	0.83	17.36
DC6	*Long − run momentum*	0.16	1.75	0.33	22.64
Panel C: Other clusters					
IC3	*Value*	0.22	1.62	0.47	20.03
IC4	*Profitability*	0.17	1.07	0.55	11.61
IC5	*Intangibles*	0.13	1.22	0.36	11.42
IC6	*Investment*	0.10	1.09	0.31	18.20
DC7	*Profitability*	0.28	1.24	0.77	12.65
DC8	*Growth*	0.09	1.10	0.27	23.37
DC9	*Accrual*	0.07	1.19	0.21	24.37

Table 16: Performance of OS maximal Sharpe ratio portfolios with NYSE breakpoints

The table reports performance of OS maximal Sharpe ratio portfolios for DC-Model and factor models, including IC-Model, FF3, Car4, Q4, FF5, and cross-sectional stock return prediction method, including Lasso, Ridge and Enet. Portfolios are constructed with NYSE breakpoints. For 6 factor models, The returns of OS maximal Sharpe ratio portfolios are constructed on a purely out-of-sample basis by using the mean and covariance matrix of estimated factors through t and tracking the post-formation t + 1 return. For cross-sectional return prediction method, the returns of portfolios are constructed by using prediction through t, where hyper-parameters are chosen with cross-validation, and tracking the post-formation t + 1 return. In the portfolios, stocks with return prediction higher than 50% percentile of return prediction in NYSE sample are in the long position and the rest are in the short position. Performance measures include sample mean, sample standard deviation, annualized Sharpe ratio and maximum drawdown (MDD) of monthly return. J is number of factors in each model. The sample period is from Jan. 1990 through Dec. 2021.

Models	K	Performance of OS maximal Sharpe ratio portfolio			
		Mean(%)	S.D.(%)	Sharpe	MDD(%)
DC	9	0.56	1.43	1.36	8.08
IC	6	0.44	1.54	0.99	18.34
FF3	3	0.39	2.42	0.56	37.64
Car4	4	0.43	2.46	0.60	32.03
Q4	4	0.30	1.62	0.64	24.62
FF5	5	0.38	1.47	0.91	20.15
Lasso	1	0.37	3.60	0.35	28.73
Ridge	1	0.43	4.02	0.37	36.21
Enet	1	0.38	3.59	0.36	27.98

Table 17: BS-CZZ method for models with NYSE breakpoints: posterior probability versus prior expectation

The table applies the BS-CZZ method and compares DC-Model with benchmarks with NYSE breakpoints. It reports posterior probability $Pr(M_l|D)$ versus prior expectation κ, which is defined in equation (24). Panel A compares the DC model with each of benchmarks and reports posterior probability that the DC model holds. Panel B compares all models simultaneously and reports posterior probability that each model holds. The sample period is from Jan. 1990 through Dec. 2021.

	$\kappa = 2.0$	2.2	2.4	2.6	2.8	3.0	3.2	3.4	3.6	3.8	4.0
Panel A: Compare the DC model with each of benchmarks											
IC	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
FF3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Car4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Q4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
FF5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Panel B: Compare the DC model with benchmarks simultaneously											
DC	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
IC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FF3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Car4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FF5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00