How is Fund Investment Exposed to Stock-level Characteristics ? Evidence from U.S. Equity Market *

Yaohan Chen [a], Xiaobin Liu [b], and Tao Zeng [b]

[a] School of Economics, Singapore Management University
[b] School of Economics, Zhejiang University

Last version: February 25, 2021

This version: February 28, 2022

Comments are welcome, currently not for circulation

Abstract

This paper documents how Instumented Principal Component Analysis (IPCA) is applied in uncovering the driving factors associated with firm-level characteristics to which fund managers holding these stocks are exposed. IPCA is a specific statistical learning methodology featuring in both latent factor structure and dynamic factor loading which accordingly is able to simultaneously handle dimensionality and time-varying parameter concern for financial econometric modelling. Linear structure is retained and therefore corresponding statistical hypothesis testing is possible to be implemented. In this paper, we firstly construct fund-level index as the measure of exposure of each fund to firm-level characteristics (commonly referred to "anomalies" in accounting or finance literature) and document the empirically stylized facts revealed from the constructed dataset. With the constructed fund-level index, we apply the novel IPCA methodology along with our proposed ℓ_1/ℓ_q-regularized IPCA to discuss how mutual fund return is exposed to characteristics of managed assets (specifically those firm-level characteristics of assets hold by fund).

Keywords: Asset Pricing; Mutual Fund; Machine-learning; PCA; ℓ_1/ℓ_q-regularization

* I am grateful for the discussion and kindly shared literatures provided by Associate Professor Tao Zeng from Zhejiang University and Assistant Professor Xiaobin Liu from Zhejiang University. All errors are my own. You may contact me at yaohan.chen.2017@phdecons.smu.edu.sg.

1 Introduction

Factor structure imposed on financial asset returns has become the pivotal part of modern financial modelling ever since the pioneering blueprint established in Fama and French (1992, 1993, 1996) and most recently the 5-factor structure proposed in Fama and French (2015). Essentially speaking, the driving force, and to some extent the most critical one, for the factor model and later many other following advanced methodologies is about handling question that how financial assets are exposed to cross-sectional information. Generally there are two kinds of methodologies dealing factor modelling in extant literature: (i) The first one corresponds to pre-specifying factors based on ex-ante established knowledge about cross-sectional accounting information. In research of this style, many benchmark factors have been established for explaining the cross-sectional variances associated with asset returns and among which the most representative ones include Fama and French (1993), Fama and French (2015), Hou, Xue, and Zhang (2015) and many other anomaly-mining related studies comprehensively documented in extant empirical finance literature (see Hou, Xue, and Zhang, 2018; Chen and Zimmermann, 2020, for relatively comprehensive summary); (ii) The second one refers to modelling factors as latent variables and applying statistical factor analysis techniques such as Principal Component Analysis (PCA) to estimate factors and factor loadings simultaneously. Studies of this kind can be at least traced back to Connor and Korajczyk (1986), Chamberlain and Rothschild (1983) and recently have been extended in a series of work in (Kozak, Nagel, and Santosh, 2018, 2020; Kozak, 2020; Lettu and Pelger, 2020a,b) by incorporating related machine-learning methodologies. But one major shortcoming inherently coming up with this standard latent factor analysis is that it does not suite the conditional asset pricing specification which relates the future asset returns with current information. Recently there are some discussions on incorporating dynamic feature in factor loading such as the work but not limited to Kelly, Pruitt, and Su (2019, 2020) and Haddad, Kozak, and Santosh (2020).

However, all these existing studies currently focus on individual stock-related assets and rarely is there any discussion investigating how fund investment is exposed to cross-sectional information by applying these recently proposed advanced methodologies. It is well-known that funds are essentially portfolios in the investment pool of fund managers, several studies have been documented in literature (see the representative ones such as Kosowski, Timmermann, Wermers, and White, 2006; Fama and French, 2010; Harvey and Liu, 2020) where the major theme of these studies is about carefully designing bootstrap procedure for inferring whether mutual funds outperform. However all these studies are all built upon reduced form regression framework with regressors, on the R.H.S. of regression equation, specified as benchmark pre-specifying factors (for instance CAPM, three-factor and four-factor benchmarks), which cannot be directly applied in the setting where funds are exposed to much more firm-level characteristics that have been comprehensively documented recently in literature (see some benchmark date sets that have recently used in Green, Hand, and Zhang, 2017; Gu, Kelly, and Xiu, 2019; Demiguel, Martín, Nogales, and Uppal, 2020; Chen, 2019; Harvey and Liu, 2014, 2015; Harvey, Liu, and Zhu, 2016; Freybergerk, Neuhierl, and Weber, 2019; Kozak,

Nagel, and Santosh, 2020; Kozak, 2020, and the most recent comprehensive summary in Chen and Zimmermann (2020)). Given this increasing availability of cross-sectional accounting information, we postulate that IPCA (Instrumented Principal Component Analysis) established in (Kelly, Pruitt, and Su, 2019, henthforth KPS2019) and (Kelly, Pruitt, and Su, 2020, henceforth KPS2020) is relatively more suitable framework within which we may conduct our empirical studies with the target as investigating how fund investment is exposed these asset characteristics at firm-level. Recently there is progress on this strand using cross-sectional information of holding assets to study investment behaviour of fund managers: Li and Rossi (2021) adds to literature from the perspective of investigating how fund investment is exposed to the characteristics of holding assets using machine-learning methodology and demonstrates that although machine-learning method may generate significant predictability, the exposure of fund to characteristics of holding assets are time-varying. Our paper is quite similar to Li and Rossi (2021) in this regard but focus more on investigating factor-exposure structure of funds using within the framework of IPCA.

Since the time when IPCA was proposed, it has achieved empirical success in some recent studies (see Kelly, Palhares, and Pruitt, 2020; Kelly, Moskowitz, and Pruitt, 2021). The success of the application of IPCA may be attributed to the following reasons: (i) linear structure is retained for IPCA and accordingly enable it to be reconciled with existing factor analysis in economics and finance (see Geweke, 1977; Sargent and Sims, 1977; Bai and Ng, 2002; Bai, 2003; Bai and Ng, 2013; Fan, Liao, and Wang, 2016; Stock and Watson, 2002) and easy to be implemented; (ii) Statistical testing is easy to be established based on bootstrap for making inference; (iii) The way linking factor loading with firm-level characteristics suggested by IPCA makes it automatically accommodate dynamic factor loading, which is one of the long-existing concern in factor modelling.

In this paper, we exploit IPCA as the major workhorse to conduct our analysis along with one novel extension of IPCA by incorporating statistical learning regularization for addressing the joint selection. This extension to some extent serve one alternative supplement to IPCA when applying testing framework established along with IPCA in determining how fund is exposed to firm-level characteristics of holding assets in that there are always some subtle issues associated with bootstrap design and bootstrap procedure is usually computationally heavy in the sense that generally there should be a reasonably large bootstrap sample sizes for obtaining reliable results. Besides, the data set of fund-level index constructed as the measure of exposure of fund to characteristics of holding assets following Kacperczyk, Sialm, and Zheng (2006) and Hoberg, Kumar, and Prabhala (2017) also to some extent facilitate the further studies corresponding to applying statistical learning or machine-learning methods in fund-relates researches. Related studies akin to this research style include Büchner (2021), in which a specific IPCA of augmented dimension is discussed so as to accommodate heterogeneity over institutional investors and this extended IPCA methodology refers to three-dimensional instrumented principal component methodology (IPCA3D).

The rest of this paper is structured as follows: In Section 2 we will discuss the basic modelling framework established on top of Instrumented Principal Component Analysis (IPCA) and the associated testing framework that serves as the workhorse of our empirical study. In Section 3 we

proceed to discuss how our fund data and firm-level characteristics are constructed and collected along with the discussion corresponding the novel fun-level index constructed as the measure of exposure of each fund to the cross-sectional information of equity assets it holds. Late in the following discussion contained in Section 4, we firstly summarize some basic information from our constructed fund-level index data for the purpose of checking the stylized fact corresponding the exposure of fund to some major anomalies both at overall level (i.e. simultaneously accommodating time-series and cross-sectional dimension) and over cross-sectional and time-series dimension respectively. Then we apply IPCA to conduct our empirical implementation. Finally Section 5 concludes this paper.

2 Basic Modelling Framework: Recap

Cross-sectional asset pricing literature postulates that returns associated with different assets are exposed to the corresponding asset characteristics, essentially different portfolios can be constructed from sorting on these characteristics and accordingly the induced returns of these constructed portfolios are commonly referred to anomalies in literature. Currently there is a vast amount of literature discussing how the cross-sectional characteristics are related with asset return ever since 3-factor modeling framework established in Fama and French (1992, 1993), in which "Size" (measured by market equity) and "Value" (measured by book-to-market ratio) along with the CAPM implied market factor are proposed as driving characteristics for stock return. Recently there is growing literature discussing how the rising machine-learning algorithms are able to play the role in identifying driving cross-sectional characteristics for different assets such as stocks (Gu, Kelly, and Xiu, 2019) and bonds (Kelly, Palhares, and Pruitt, 2020), however by far rarely is there any discussion corresponding to the fund-related returns following this strand.

The way that distinguishes IPCA from conventional factor modeling framework, as to be discussed in the preceding discussion, is the parametrization of factor beta as to link the firm and and asset characteristics. To facilitate the following discussion, we may firstly extend IPCA from KPS2019 and KPS2020 to fit this setting as following

$$r_{i,t+1} = \alpha_{i,t} + \beta_{i,t} f_{t+1} + \epsilon_{i,t+1} \tag{1}$$

$$\alpha_{i,t} = z_{i,t}^{\top} \Gamma_{\alpha} + \nu_{\alpha,i,t}, \quad \beta_{i,t} = z_{i,t}^{\top} \Gamma_{\beta} + \nu_{\beta,i,t} \tag{2}$$

where $z_{i,t}$ is a $L \times 1$ vector collecting L normalized characteristics while Γ_{α} and Γ_{β} are $L \times 1$ vector and $L \times K$ matrix respectively, similarly referred to factor loadings in standard literature. Accordingly $\alpha_{i,t}$ and $\beta_{i,t}$ are 1×1 scalar and $1 \times K$ row vector respectively. K denotes the specified number of factors. [1] Both Γ_{α} and Γ_{β} are identified only up to rotation and this could be alternatively interpreted as that K-factor IPCA finds the K-dimensional space spanned by estimated factors but the rotation of these factors does not change the corresponding model fit. The identification scheme is same as the way suggested in Kelly, Pruitt, and Su (2020) with additional normalization imposed

[1] For the sake of matrix dimension match, we have to assume $\nu_{\alpha,i,t}$ and $\nu_{\beta,i,t}$ as scalar and $1 \times K$ row vector as well.

such that $\Gamma_\beta^\top \Gamma_\beta$ is identity matrix. As suggested in (1) and (2), we are interested in the following to be discussed null-hypothesis testing in the following discussion we will demonstrate that suggested framework conducting null-alternative hypothesis testing does not suffer from the identification issue of the specified factor structure. We may plug (2) into (1), which yields

$$r_{i,t+1} = z_{i,t}^\top \Gamma_\alpha + z_{i,t}^\top \Gamma_\beta f_{t+1} + \epsilon_{i,t+1}^*, \qquad \epsilon_{i,t+1}^* = \epsilon_{i,t+1} + \nu_{\alpha,i,t} + \nu_{\beta,i,t} f_{t+1}. \tag{3}$$

Several advantaged of IPCA is brought with this advantage as our target is to obtain estimation of $\Gamma = \left[\Gamma_\alpha, \Gamma_\beta\right]$ and factors collected in $\{f_t\}$ in that alternating least squares (ALS) methodology can be applied alternatively between minimizing Γ while holding $\{f_t\}$ fixed, and minimizing over f_t while holding Γ fixed. This handles unbalanced panels as easily as pooled OLS. More comprehensive technical details and discussions are well documented in KPS2020, which is one of the most recent literature the readers may refer to. We will briefly revisit ALS ideas and statistical testing framework established in KPS2020 in this section before we move on to the empirical application part as all these established estimation and testing procedure serve as the workhorse though which our empirical study is conducted and lay the foundation for our proposed ℓ_1/ℓ_q-regularized IPCA for joint selection that is to be covered in subsection 4.3.

2.1 Estimation

Note that we may characterize reduced form of IPCA equation (3) in compact matrix form as following

$$\mathbf{r}_{t+1} = \mathbf{Z}_t \Gamma_\beta f_{t+1} + \boldsymbol{\epsilon}_{t+1}^* \tag{4}$$

where the additionally introduced notations are summarized as following

$$\begin{aligned}
\mathbf{Z}_t &: \ N \times L \text{ matrix, normalized firm-level characteristics.} \\
\mathbf{r}_{t+1} &: \ N \times 1 \text{ vector, collecting individual asset returns at } t+1. \\
\boldsymbol{\epsilon}_{t+1}^* &: \ \text{Vector collecting } \epsilon_{i,t+1}^*, \text{ where } \epsilon_{i,t+1}^* = \epsilon_{i,t+1} + \nu_{\alpha,i,t} + \nu_{\beta,i,t} f_{t+1}.
\end{aligned}$$

From the least squares perspective, the objective of IPCA in terms of estimation is to obtain estimation of Γ_β and $\{f_{t+1}\}$ jointly through the following quadratic optimization problem.

$$\min_{\Gamma_\beta, \{f_{t+1}\}} \frac{1}{T} \sum_{t=1}^{T-1} \left(\mathbf{r}_{t+1} - \mathbf{Z}_t \Gamma_\beta f_{t+1}\right)^\top \left(\mathbf{r}_{t+1} - \mathbf{Z}_t \Gamma_\beta f_{t+1}\right) \tag{5}$$

For the sake of retaining focus on the main ideas associated with ALS estimation of IPCA, we encompass intercept term as one factor. In other word, if first element of f_{t+1} is 1 then the first column of Γ_β for this scenario maps exactly to Γ_α. The way KPS2019 derives their formula for alternating least squares is based on rewriting this minimization as stacked linear form whereas detailed discussion is ignored in that it is not major focus of that paper, but it would be inspiring

to discuss corresponding details more explicitly here to see the potential for extension to framework for joint selection by incorporating regularization. Summation taken over time-series dimension is separable and first order conditions taken with respect to f_{t+1} with Γ_β fixed for each t. This is OLS-style F.O.C. and yields

$$\hat{f}_{t+1} = \left(\Gamma_\beta^\top \mathbf{Z}_t^\top \mathbf{Z}_t \Gamma_\beta\right)^{-1} \Gamma_\beta^\top \mathbf{Z}_t^\top \mathbf{r}_{t+1} \tag{6}$$

With fixed estimation $\left\{\hat{f}_{t+1}\right\}$, the F.O.C. implied from taking first order derivative with respect to Γ_β is given as

$$\text{vec}\left(\hat{\Gamma}_\beta^\top\right) = \left(\sum_{t=1}^{T} \mathbf{Z}_t^\top \mathbf{Z}_t \otimes f_{t+1} f_{t+1}^\top\right)^{-1} \left(\sum_{t=1}^{T-1} \left(\mathbf{Z}_t^\top \otimes \mathbf{r}_{t+1}\right) f_{t+1}\right) \tag{7}$$

In appendix A we demonstrate how it is obtained, which is not covered that much in details in KPS2019 and KPS2020. To sum up, iterating over F.O.C. characterized by (6) and (7) alternatively starting from an initialized guess of Γ_β until tolerance condition has been satisfied yields the this alternating least squares estimation of $\hat{\Gamma}_\beta$ and $\left\{\hat{f}_{t+1}\right\}$ respectively.

2.2 Testing based ALS estimation

In this section, we revisit how statistical testing framework is to be established based ALS estimation $\hat{\Gamma}_\beta$ and $\hat{f}_{t+1}$ suggested from (6) and (7).

2.2.1 Testing intercept term

This testing framework is fully characterized as following

$$H_0 \ : \ \Gamma_\alpha = \mathbf{0}_{L\times 1}$$

$$H_1 \ : \ \Gamma_\alpha \neq \mathbf{0}_{L\times 1}$$

The proposed Wald-type statistic is

$$W_\alpha = \hat{\Gamma}_\alpha^\top \hat{\Gamma}_\alpha \tag{8}$$

and based on this the corresponding inference is implemented through the following bootstrap procedure. To facilitate our discussion of bootstrap procedure, we characterize (1) in matrix form as following [2]

$$\mathbf{x}_t = \mathbf{Z}_t^\top \mathbf{r}_{t+1} = \underbrace{\left(\mathbf{Z}_t^\top \mathbf{Z}_t\right)}_{\mathbf{w}_t} \Gamma_\alpha + \left(\mathbf{Z}_t^\top \mathbf{Z}_t\right) \Gamma_\beta f_{t+1} + \mathbf{Z}_t^\top \boldsymbol{\epsilon}_{t+1}^* \tag{9}$$

where all the notations in bold-face refer to vectors or matrices collecting cross-sectional elements associated with timing index t and $t+1$. Specifically, we introduce another notation $\mathbf{x}_t$, which is as following

[2] Alternatively, it is possible to interpret (9) in this way such that managed portfolio returns are exposed to common factors through factor loading $\left(\mathbf{Z}_t^\top \mathbf{Z}_t\right)\Gamma_\beta$.

$\mathbf{x}_t$: $L \times 1$ vector, "managed portfolio" (KPS2019).

Remark 2.1 *Although it seems that (9) suggests a balanced matrix manipulation, we want to emphasize it here that the real data exploited for empirical analysis is essentially an unbalanced panel data. Accordingly for each element $x_{l,t}$ indexed by pair (l,t) in $\mathbf{x}_t$, it is constructed as the dot product of sub-vector of the l-th column of $\mathbf{Z}_t$, for which the indices associated with elements contained in these vectors also map to non-missing observations in $\mathbf{r}_{t+1}$. In other words, for each fixed t, we eliminate missing elements either for the L columns of $\mathbf{Z}_t$ or $\mathbf{r}_{t+1}$.*

The bootstrap procedure is implemented as following

Step 1. Estimating unrestricted model without imposing restrictions $\Gamma_\alpha = 0$, thus we allow intercept in the modelling and retain the corresponding estimation:

$$\hat{\Gamma}_\alpha,\ \hat{\Gamma}_\beta,\ \{f_t\}_{t=1}^T$$

Step 2. Next for $b = 1, \ldots, B$, generating the b-th bootstrap sample as

$$\tilde{\mathbf{x}}_{t+1}^b = \left(\mathbf{Z}_t^\top \mathbf{Z}_t\right) \hat{\Gamma}_\beta \hat{f}_{t+1} + \tilde{\mathbf{d}}_{t+1}^b, \qquad \tilde{\mathbf{d}}_{t+1}^b = q_{1,t+1}^b \hat{\mathbf{d}}_{q_{2,t+1}^b} \tag{10}$$

where $\hat{\mathbf{d}}_{q_{2,t+1}^b}$ refers to the $q_{2,t+1}^b$-th element extracted from $\left\{\hat{\mathbf{d}}_t\right\}_{t=1}^T$ and each $\hat{\mathbf{d}}_t$ serves as the residual counterpart of $\mathbf{Z}_t^\top \boldsymbol{\epsilon}_{t+1}^*$ implied from (9); and $q_{1,t+1}^b$, as suggested in KPS2019, refers to random variables sampled from Student-t distribution with unit variance and five degrees of freedom. Then for each b, using the bootstrap sample to re-estimate the unrestricted model and retain the corresponding estimated test statistic as

$$\tilde{W}_\alpha^b = \tilde{\Gamma}_\alpha^{b\top} \tilde{\Gamma}_\alpha^b \tag{11}$$

Step 3. p-value associated this null-hypothesis testing is implied from the empirical null distribution generated from bootstrap sample by specifying it as the fraction of bootstrapped statistics $\tilde{W}_\alpha$ that exceeds the estimated W_α from data. [3]

2.2.2 Testing instrument significance

The basic testing procedure methodologically inherits much from the testing framework established for testing significance of intercept term. A specific test designed for testing whether corresponding

[3] This step could be intuitively interpreted as following: the bootstrapped data $\tilde{\mathbf{x}}_{t+1}^b$ is re-sampled data for given $\mathbf{Z}_t$, $\hat{\Gamma}_\beta$ and $\hat{f}_{t+1}$ with restrictions imposed such that $\Gamma_\alpha = 0$ and accordingly bootstrapped data $\tilde{\mathbf{x}}_{t+1}^b$ generated via re-sampling residuals reveals information associated with the imposed restrictions. $\tilde{\Gamma}_\alpha^b$ is unrestricted estimation of Γ_α associated with bootstrapped data, which naturally induces empirical null distribution. Alternatively, the generated p-value claimed in **Step 3.** implies whether unrestricted estimation Γ_α is significantly large in terms of Euclidean norm with respect to the bootstrapped empirical null distribution associated with $\tilde{\Gamma}_\alpha^b$.

instrument (firm-level characteristic) significantly contribute to $\beta_{i,t}$. Thus we are interested in the partition of the loading matrix as following

$$\Gamma_\beta = \left[\gamma_{\beta,1}, \ldots, \gamma_{\beta,L}\right]^\top \tag{12}$$

The proposed test-statistics are based on the following null hypothesis and alternative hypothesis.

$$\begin{aligned} H_0 \ &: \ \Gamma_\beta = \left[\gamma_{\beta,1}, \ldots, \gamma_{\beta,l-1}, \mathbf{0}_{K\times 1}, \gamma_{\beta,l+1}, \ldots, \gamma_{\beta,L}\right]^\top \\ H_1 \ &: \ \Gamma_\beta = \left[\gamma_{\beta,1}, \ldots, \gamma_{\beta,L}\right]^\top \end{aligned}$$

with the induced Wald-type statistic constructed as following

$$W_{\beta,l} = \hat{\gamma}_{\beta,l}^\top \hat{\gamma}_{\beta,l} \tag{13}$$

The bootstrap procedure implemented for the above specified hypothesis-testing framework is in analogy to the one described previously for testing Γ_α, for comparison purpose we summarize the corresponding steps in the following

Step 1. Estimating unrestricted model without imposing restriction $\gamma_{\beta,l} = \mathbf{0}_{K\times 1}$ so as to obtain estimation of Γ_β such that

$$\hat{\Gamma}_\beta = \left[\hat{\gamma}_{\beta,1}, \ldots, \hat{\gamma}_{\beta,l-1}, \gamma_{\beta,l}, \hat{\gamma}_{\beta,l+1}, \ldots \hat{\gamma}_{\beta,L}\right]^\top \tag{14}$$

Step 2. Next for $b = 1, \ldots, B$, generating the b-th bootstrap sample as

$$\tilde{\mathbf{x}}_{t+1}^b = \hat{\Gamma}_\alpha + \left(\mathbf{Z}_t^\top \mathbf{Z}_t\right) \tilde{\Gamma}_\beta \hat{f}_{t+1} + \tilde{\mathbf{d}}_{t+1}^b \tag{15}$$

where

$$\tilde{\Gamma}_\beta = \left[\hat{\gamma}_{\beta,1}, \ldots, \hat{\gamma}_{\beta,l-1}, \mathbf{0}_{K\times 1}, \hat{\gamma}_{\beta,l+1}, \ldots, \hat{\gamma}_{\beta,L}\right]^\top \tag{16}$$

and $\tilde{\mathbf{d}}_{t+1}^b$ denotes the bootstrapped sample of $\left\{\hat{\mathbf{d}}_t\right\}$ with each $\hat{\mathbf{d}}_t$ indicating the residual counterpart of $\mathbf{Z}_t^\top \boldsymbol{\epsilon}_{t+1}^*$ implied from unrestricted estimation. Then bootstrapped sample $\hat{\mathbf{x}}_t^b$ is applied to re-estimate the alternative model (i.e. unrestricted model) and obtain estimation of l-th column of $\tilde{\Gamma}_\beta$, denoted by $\tilde{\gamma}_{\beta,l}$. Accordingly the bootstrapped test statistic is given as

$$\tilde{W}_{\beta,l}^b = \tilde{\gamma}_{\beta,l}^\top \tilde{\gamma}_{\beta,l} \tag{17}$$

Step 3. p-value associated this null-hypothesis testing is implied from the empirical null distribution generated from bootstrap sample by specifying it as the fraction of bootstrapped statistics $\tilde{W}_{\beta,l}$ that exceeds the estimated $W_{\beta,l}$ from data.

2.2.3 Testing additional observable factors significance

Likewise, we may augment factor space by adding observable factors documented in literature (for instance, Fama-French 3-factors) and construct Wald-type statistic to test funds are exposed to these observable factors. We temporarily focus on the restricted model (i.e. $\Gamma_\alpha = \mathbf{0}$) but it would be convenient to extend by including intercept term. Specifically we augment factor space as $\tilde{f}_{t+1} = \left[f_{t+1}^\top, g_{t+1}^\top\right]^\top$ and correspondingly $\tilde{\Gamma} \equiv \left[\Gamma_\beta, \Gamma_\delta\right]$, where g_{t+1} denotes the $M \times 1$ vector collecting added observable factors and Γ_δ as $L \times M$ matrix denotes the accompanied factor loading matrix. This specification obviously implies that each asset return is exposed to common factors in the following way

$$r_{i,t+1} = z_{i,t}^\top \Gamma_\beta f_{t+1} + z_{i,t}^\top \Gamma_\delta g_{t+1} + \epsilon_{i,t+1} + \nu_{\beta,i,t} f_{t+1} + \nu_{\delta,i,t} g_{t+1} \tag{18}$$

where we have implicitly imposed the structure that $\delta_{i,t} = z_{i,t}^\top \Gamma_\delta + \nu_{\delta,i,t}$ and likewise we assume that $\delta_{i,t}$ and $\nu_{\delta,i,t}$ as $1 \times M$ row vector to make corresponding matrix computation reconcilable. Then we can apply previously established framework almost in the same way here to construct Wald-type statistic for checking whether Γ_δ is significantly away from zero. Thus for this scenario the corresponding null and alternative hypothesis are claimed respectively as following

$$H_0 \ : \ \Gamma_\delta = \mathbf{0}_{L \times M}$$

$$H_1 \ : \ \Gamma_\delta \neq \mathbf{0}_{L \times M}$$

the Wald-type statistic is constructed $W_\delta = \text{vec}\left(\hat{\Gamma}_\delta\right)^\top \text{vec}\left(\hat{\Gamma}_\delta\right)$, where $\hat{\Gamma}_\delta$ denotes the unrestricted estimation of Γ_δ in (18). Wild bootstrap procedure can be applied similarly here via resampling residuals from $\mathbf{Z}_t^\top \mathbf{r}_{t+1} - \mathbf{Z}_t^\top \hat{\Gamma}_\beta f_{t+1} - \mathbf{Z}_t^\top \hat{\Gamma}_\delta g_{t+1}$. Then use the bootstrapped data to obtain unrestricted estimation of Γ_δ denoted by $\tilde{\hat{\Gamma}}_\delta$ and accordingly the bootstrapped Wald-statistic $\tilde{\hat{W}}_\delta = \text{vec}\left(\tilde{\hat{\Gamma}}_\delta\right)^\top \text{vec}\left(\tilde{\hat{\Gamma}}_\delta\right)$. Finally the p-value associated with this hypothesis testing is calculated as the fraction of bootstrapped $\tilde{\hat{W}}_\delta$ that exceeds W_δ.

3 Data

3.1 Fund holding data

We follow the standard procedure as in Kacperczyk, Sialm, and Zheng (2006) and Hoberg, Kumar, and Prabhala (2017) to collect, clean and construct fund-related data from the Center for Research in Security Prices (CRSP) Survivorship Bias Free Mutual Fund Database and later merge it with Thompson Financial CDA/Spectrum holdings database and CRSP stock price data following methodology of Kacperczyk, Sialm, and Zheng (2005). Furthermore to shrink the universe of funds, we leverage the methodology suggested in Kacperczyk, Sialm, and Zheng (2006), which is essentially a sequential algorithm. Thus we firstly select funds whose Lipper Classification Code (identified by `lipper_class`) is one of the following: EIEI, LCCE, LCGE, LCVE, MCCE, MCGE, MCVE,

MLCE, MLGE, MLVE, SCCE, SCGE, or SCVE; If the Lipper classification code is missing, funds are selected as those whith "Straight Insights" Objective code (identified by `si_obj_cd`) as one of the following: AGG, GMC, GRI, GRO, ING, or SCG; Then if both codes are missing, funds are selected as those with Wiesenberger objective codes (identified by `wbrger_obj_cd`) as one of the following: G, G-I, GCI, LTG, MCG, or SCG, or those with "Policy" code of CS. Besides, the universe of funds are restricted to those funds whose lifetime average investment in equity is at least 80% and those funds holding fewer than 10 stocks and managed assets less than $5 million are excluded from the fund universe as well. Portfolio holdings are matched to mutual funds using `MFLINK` tables developed by Russ Wermers and made available via Wharton Research Data Services.

3.2 Firm-level characteristic data

Firm-level characteristic data used for empirical analysis in this paper follows the strand of literation either from accounting or finance including (Green, Hand, and Zhang, 2017; Gu, Kelly, and Xiu, 2019; Demiguel, Martín, Nogales, and Uppal, 2020; Chen, 2019; Harvey and Liu, 2014, 2015; Harvey, Liu, and Zhu, 2016; Freybergerk, Neuhierl, and Weber, 2019; Kozak, Nagel, and Santosh, 2020; Kozak, 2020). Several standard databases have been available for researches based on this work, but to the limited knowledge of us, the most recent work done in Chen and Zimmermann (2020) (henceforth CZ2020a) is by far the most recent and comprehensive one that has successfully covered almost all the major documented anomalies in literature [4]. The work done by CZ2020a is relatively a successful response to the call for transparency and cooperation and in empirical finance research, which is claimed in Welch (2019). We focus on 208 firm-level characteristics constructed in CZ2020a including "Size" (measured as the market equity value associated with each individual stock) characteristic constructed following standard procedure as in Fama and French (1992, 1993, 1996, 1997). We summarize the basic information of these 208 characteristics in Table D.1 as readers' references for details.

For the constructed firm-level characteristics, each characteristic is normalized crosss-sectionally to make it lie in between 0 and 1 in a way such that

$$rc_{i,t}^{s} = \frac{\text{rank}\left(c_{i,t}^{s}\right)}{n_t + 1} \tag{19}$$

where $c_{i,t}^{s}$ refers to the originally unscaled firm-level characteristic (indexed by superscript s) associated with stock i at time t and n_t refers to the total number of firms available for observations at time t. $\text{rank}(\cdot)$ denotes the cross-sectional ranking order of specific variable.

[4] We acknowledge the codes and data kindly shared by the authors and their efforts on constantly maintaining and updating the data. Both the codes and data are available at the authors' maintained website https://sites.google.com/site/chenandrewy/open-source-ap?authuser=0. The dataset used in this paper mainly corresponds to the released version "Version 0.1.2, Patch, July 23, 2020". Currently there is a newly released version, "March 2021 Data Release: Major Update".

3.3 Fund-level characteristic data

In this section, we briefly discuss how we link the fund-level characteristic data and firm-level characteristic data discussed in subsection 3.1 and subsection 3.3. Specifically, we merge the fund data (including `"date"` at monthly frequency and fund identifier `"fundno"`) with collected firm-level characteristic data using `"date"` and `"permno"` as key. Once these two datasets are merged as the way suggested, we define the fund-level characteristic data as the index measuring the extent to which each fund is exposed to the corresponding characteristics. Specifically, we define fund-level index as weighted average of normalized firm-level characteristics as following [5]

$$z^s_{j,t} = \sum_{i=1}^{n^s_{j,t+1}} w^s_{j,i,t} \cdot rc^s_{i,t} \tag{20}$$

where notations involved in (20) are little bit complicated and accordingly the corresponding detailed notation explanations are summarized as following,

$z^s_{j,t}$: fund level index, associated with fund j at time t, exposed to characteristic s.

$n^s_{j,t+1}$: total number of stocks hold by fund j at time $t+1$, that are available for being as the observations at time t.

$w^s_{j,i,t}$: weights used for aggregating firm-level characteristics. Thus the weight assigned to stock i hold by fund j, the superscript s denote the normalized characteristic to which fund j is exposed.

$rc^s_{i,t}$: normalized characteristic s at firm-level associated with stock i at time t as expressed in (19).

More specifically, the weights adopted for constructing (20) is constructed from lagged holding value, denoted by $\text{HVALUE}_{j,i,t}$, which is calculated as

$$\text{HVALUE}_{j,i,t} = \texttt{shares}_{j,i,t+1} \times \texttt{cfacshr}_{j,i,t+1} \times \left(\frac{|\texttt{prc}|}{\texttt{cfacpr}}\right)_t \tag{21}$$

where the corresponding WRDS identifier acronyms are explained as following respectively

`shares` : Shares held by fund at the end of each quarter.

`cfacshr` : Cumulative Factor to Adjust Shares/Vol.

`cfacpr` : Cumulative Factor to Adjust Prices.

`prc` : Price or Bid/Ask Average. ($|\cdot|$ refers to evaluated absolute values).

and subscript i refers to the index of stocks hold by fund. Accordingly, for each fund j and firm-level characteristic s, $w^s_{j,i,t}$ is self-normalized $\text{HVALUE}_{j,i,t}$.

[5] This constructed fund-level index is akin to the Anomalies Investing Measure (AIM) proposed in Ali, Chen, Yao, and Yu (2008).

Remark 3.1 $\texttt{shares}_{j,i,t+1}$ *refers to shares of stock i hold by fund j at time $t+1$, but when it is time to consider how to merge back firm-level characteristic with fund holding data, we have to take the lagged information into account. Alternatively speaking, we have to exploit lagged holding values as weights to rescale firm-level characteristics as to construct fun-level index measuring exposure of fund to each characteristic, and accordingly lagged cross-sectional information of firm-level characteristics has to be accommodated for alignment as well. One of the similar discussion has been done in (Lettau, Ludvigson, and Manoel, 2021, henceforth LLM2021) as well, but the discussion there does not cover the application of IPCA methodology to accommodate dynamic exposure of fund return to factor structure that summarizes cross-sectional information; Besides, the objective of LLM2021 is slightly different from ours as well. However, some of the methodologies exploited are essentially related and accordingly some of empirical findings coincide with those claimed in LLM2021. For instance, we also find that those actively managed mutual funds in U.S. equity market generally demonstrate tilted distribution in terms of the distribution over value-related measure such as the conventional book-to-market ratio.* [6] *Although LLM2021 focus primarily on holdings of mutual funds, one of the reason they claim for using holdings as the description of mutual fund strategies rather than factor loadings is factor loadings naturally vary over time and accordingly are hard to handle.*

Besides, we also include some standard fund-level characteristics such as fund flow sand fund age in our constructed dataset. For the fund flow, we follow the way as in Barber, Huang, and Odean (2016) through which fund flow for fund j at the end of month t is calculated as the percentage growth of new funds at the the end of each month such that

$$\texttt{flow}_{j,t} = \frac{\texttt{TNA}_{j,t}}{\texttt{TNA}_{j,t-1}} - (1 + r_{j,t}) \tag{22}$$

where $\texttt{TNA}_{j,t}$ refers to the total net assets under management of fund j. While fund age associated with specific fund is simply the number of months from the inception of fund.

4 Empirical Findings

In this section, we proceed to the discussion corresponding to how fund managers' investment behaviour (characterized by returns associated with each fund) is exposed to the constructed fund-level index aggregating information contained in assets hold by each fund. We collect mutual fund return data at monthly frequency from `CRSP.Monthly_Returns` database and require constructed data to be identifiable from `MFLINKS` in a a specific way such that `crsp_fundno` as the key is identifiable from `MFL.MFLINK1` database and those observations with missing `wficn` key are eliminated. [7]

[6] One of the finding in LLM2021 is that BM ratio of mutual fund is tilted towards low BM value rather than high BM ratios.

[7] As our construction of fund holding data requires funds to be identifiable from `MFL.MFLINK2` database via `fundno` key and for each cross-section sliced from this constructed panel data, thus on average we have around 100 funds at sliced cross-section for each fixed t.

4.1 Summary

It would be interesting to check how funds collected in the universe of funds investigated are distributed on specific characteristics, thus `AM` (Total Assets to Market), `BM` (Book to market using most recent ME), `Mom12m` (12-month Momentum), `Mom6m` (6-month Momentum), `Mom1m` (Short term reversal), `Size` (Size measured market equity value).

[Place Figure 1 about here]

To see how the cross-sectional median varies over time, we plot cross-sectional median of constructed fund-level indices exposed to the above mentioned anomalies (firm-level characteristics) as following

[Place Figure 2 about here]

It would also be interesting to make a comparison about how funds are distributed on characteristics associated with intangible information (for instance, Daniel and Titman, 2006; Eisfeldt, Kim, and Papanikolaou, 2020) [8]. Following the same procedure as implemented in the preceding discussion, we demonstrate how funds are distributed over intangible returns associated with fundamental-price ratios documented in Daniel and Titman (2006, henceforth DT2006). Details corresponding to how intangible returns are established in connection with fundamental-price ratios are well documented in DT2006 but I will briefly discuss how it is established in the note attached to the following figure.

[Place Figure 3 about here]

Similarly, we plot cross-sectional median of constructed fund-level indices exposed to the above mentioned anomalies associated with intangible returns discussed in DT2006 as following

[Place Figure 4 about here]

Since one of the major target of this paper is to unravel how fund managers' decisions (characterized by return associated with each fund) is connected with our constructed fund-level index measuring the extent to which each fund is exposed to the corresponding firm-level characteristics, for the sake of data completeness, we retain our focus on the selected 16 characteristics from 208 firm-level characteristics, which essentially excludes those "score"-like characteristics taking discrete values. [9]

---

[8] Characteristics associated intangible information are initially discussed formerly in Daniel and Titman (2006), where the authors suggest a way to decompose historical return into "tangible" parts that can be unravelled solely based on the past fundamental measures and correspondingly the "intangible" parts as the past returns remaining unexplained.

[9] These 16 firm-level characteristics perhaps are the most prominent ones, it would need efforts to pin down a relatively larger universe from these 208 characteristics of CZ2020a, which is currently listed in our research agenda. However, as expansion of universe of firm-level characteristics data would become less informative as the balanced panel data which is naturally should be the input of IPCA or our extended ℓ_1/ℓ_q-regularized IPCA.

4.2 Individual testing

For this section we specifically retain our focus on 16 major anomalies `AM`, `BM`, `Mom12m`, `Mom6m`, `Mom1m`, `Size`, `ChangeRoA`, `ChangeRoE`, `IdioRisk`, `IdioVol3F`, `IdioVolAHT`, `IdioVolCAPM`, `IntanBM`, `IntanCFP`, `IntanEP`, `IntanMom`. And for testing the contribution from each of these focused anomalies, we basically implement the testing procedure discussed in subsection 2.2.2 based on the constructed Wald-type statistic, $W_{\beta,l}$, measuring the distance of associated rows of estimated Γ_β away from zero.

[Place Table 1 about here]

In the above table, we summarize bootstrapped Wald-type statistic $W_{\beta,l}$ for testing statistical significance of individual instruments (firm-level characteristics or anomalies) and the associated bootstrapped p-values via the comparison between Wald-type statistic constructed from unrestricted estimation of Γ_β and the bootstrapped sample $\left\{\tilde{W}^b_{\beta,l}\right\}_{b=1}^{B}$ with null-hypothesis testing restriction imposed on Γ_β as we have discussed in the previous section. We basically summarizes 5 different cases with different specification of factor numbers such that K ranges from 2 to 6 and for each case the column-pair documents testing statistic $W_{\beta,l}$ and bootstrapped p-value respectively. "***", "**" and "*" refers to statistical significance levels at 1%, 5% and 10% respectively.

One of the suggested result from this table is that in terms of statistical significance, "Value" (measured as our constructed fund-level index exposed to firm-level book-to-market ratio) does significantly matter for the most for the exposure of equity fund to different factors that summarizes the cross-sectional variances under different factor specifications ranging from $K = 2$ to $K = 6$. This finding to some extent suggests that fund managers (at least for equity fund managers given our filtering scheme) still pays relatively more attention to the values (measures by book-to-market ratio) as choosing assets to construct portfolios. Moreover as the number of specified factor increases such as $K = 5$ and $K = 6$, the contribution from size of assets hold by funds to the cross-sectional exposure increases and this is consistent with the intuition underlying factor modelling that as cross-sectional dimension increases more factors are needed to summarize the cross-sectional information. We also report results corresponding to fund-level characteristic universe slightly augmented to 19 characteristics with additional basic fund-level characteristics (`TNA` (total net assets under management), `flow` (fund flow) and `age`(fund age)) added. We find that these three basic fund-level characteristics significantly matters for the structure of factor exposure but the contributions to factor exposure from "value" and "size" related fund-level characteristics are still relatively significant at certain significance level. Detailed results are summarized in Table B.1.

To conduct a relatively more comprehensive analysis, we expand the universe to match the firm-level characteristics discussed in Kozak (2020) as much as possible, although to make such a kind of expansion applicable we need to replace matrix inverse with Moore-Penrose inverse instead. Thus for the following discussion we retain our focus on the fund-level indices as the exposure to the following listed firm-level characteristics of holding assets: `Size`, `BM`, `GP`, `Profitability`, `PS`, `DebtIssuance`, `ShareRepurchase`, `ShareIss1Y`, `Accruals`, `AssetGrowth`, `ChAssetTurnover`, `DivYield`, `EP`, `CF`, `NOA`, `Investment`, `InvGrowth`, `Leverage`, `SP`, `GrLTNOA`, `Mom6m`,

`IndMom`, `ShortInterest`, `Mom12m`, `Mom1m`, `Mom18m13m`, `EarningsSurprise`, `ChangeRoA`, `ChangeRoE`, `IdioRisk`, `IdioVol3F`, `IdioVolAHT`, `IdioVolCAPM`, `CompEquIss`, `CompositeDebtIssuance`, `ShareVol`, `EquityDuration` (37 characteristics in total). Detailed meaning of these characteristic acronyms are still contained in Table D.1. Besides, we add 3 basic normalized fund-level characteristics: `TNA` (total net assets under management); `flow` (fund flow); `age`(fund age) to augment the fund-level index space. Thus dimension of the finally augmented space is $37 + 3 = 40$. Surprisingly, once within this framework in which relatively more fund-level characteristics are accommodated, what remains the most significant for measuring the factor structure of fund are basic fund-level characteristics rather than the weighted aggregation of characteristics of holding assets. Detailed results are summarized in Table C.1.

4.3 Joint selection

In this section, we proceed to propose one alternative incorporating ℓ_1/ℓ_q regularized least squares in the alternating least squares procedure associated with IPCA. This proposed methodology may serve as an alternative to complement results suggested from bootstrapped IPCA. Recall the alternating least squares objective function as in (5), we may focus on the the following equivalent transformation so that we are able to impose regularization penalty for this objective function

$$\min_{\Gamma_\beta,\{f_{t+1}\}} \frac{1}{T}\sum_{t=1}^{T}\left(\mathbf{r}_{t+1}-\mathbf{Z}_t\Gamma_\beta f_{t+1}\right)^\top\left(\mathbf{r}_{t+1}-\mathbf{Z}_t\Gamma_\beta f_{t+1}\right) \quad =$$

$$\min_{\Gamma_\beta,\{f_{t+1}\}} \frac{1}{T}\sum_{t=1}^{T}\left\{\underbrace{\mathbf{r}_{t+1}}_{\substack{N\times 1\\ \text{column vector}}} - \underbrace{\left(\mathbf{Z}_t\otimes f_{t+1}^\top\right)}_{\substack{N\times(L\times K)\\ \text{Design Matrix}}}\underbrace{\text{vec}\left(\Gamma_\beta^\top\right)}_{\substack{(L\times K)\times 1\\ \text{column vector}}}\right\}^\top\bullet\left\{\underbrace{\mathbf{r}_{t+1}}_{\substack{N\times 1\\ \text{column vector}}} - \underbrace{\left(\mathbf{Z}_t\otimes f_{t+1}^\top\right)}_{\substack{N\times(L\times K)\\ \text{Design Matrix}}}\underbrace{\text{vec}\left(\Gamma_\beta^\top\right)}_{\substack{(L\times K)\times 1\\ \text{column vector}}}\right\}.$$

To facilitate our following discussion, we rewrite the least squares objective function above in a parsimonious way such that

$$\mathcal{R} = \mathcal{Z}\mathcal{V}. \tag{23}$$

where

$$\begin{aligned}
\mathcal{R} &:= \left(\mathbf{r}_1^\top,\ldots,\mathbf{r}_T^\top\right)^\top && N\times T \text{ column vector}\\
\mathcal{Z} &:= \left[\mathcal{Z}_{G_1},\ldots,\mathcal{Z}_{G_L}\right] && (N\times T)\times(L\times K) \text{ matrix}\\
\mathcal{Z}_{G_l} &:= \left[\tilde{\tilde{z}}_{1l}\otimes f_2^\top;\ldots;\tilde{\tilde{z}}_{Tl}\otimes f_{T+1}^\top\right] && (N\times T)\times K \text{ matrix}\\
\mathcal{V} &:= \text{vec}\left(\Gamma_\beta^\top\right) && L\times K \text{ column vector}
\end{aligned}$$

and $\tilde{\tilde{z}}_{tl}$ $(1 \leqslant t \leqslant T, 1 \leqslant l \leqslant L)$ refers to l-th column of $\mathbf{Z}_t$ and $\mathcal{V}_{G_l}$ $(1 \leqslant l \leqslant L)$ are non-overlapping blocks of equivalent size such that each $\mathcal{V}_{G_l}$ as $K \times 1$ vector is the transpose of l-th row of Γ_β. With this representation as in (23) established, we propose to focus on the following ℓ_1/ℓ_q regularized least squares optimization instead of the original global least squares optimization

$$\min_{\boldsymbol{\mathcal{V}}, \{f_{t+1}\}} \frac{1}{2} \|\boldsymbol{\mathcal{R}} - \boldsymbol{\mathcal{Z}}\boldsymbol{\mathcal{V}}\|_2^2 + \lambda \sum_{l=1}^{L} w_l^g \|\mathcal{V}_{G_l}\|_q \tag{24}$$

where in addition λ and $\left\{w_l^g\right\}_{l=1}^L$ are introduced to denote regularization parameters and q as the subscript indicates which norm to be used for regularization. Usually q is specified as $q = 2$ and accordingly Euclidean norm is applied for block regularization. Moreover, to make our prior knowledge has the least influence on the embedded regularized least squares selection procedure, we assume that $w_1^g =, \ldots, = w_L^g$ and this could be relaxed and easily customized via `SLEP` (**S**parse **L**earning with **E**fficient **P**rojections) developed by Liu, Ji, and Ye (2009). We may solve optimization characterized by (24) in alternating procedure similar to the original ALS. Note that for any given Γ_β (equivalently $\boldsymbol{\mathcal{V}}$ as well), the F.O.C. regarding $\{f_{t+1}\}$ suggested in (6) still applies for this regularized optimization while for fixed $\left\{\hat{f}_{t+1}\right\}$, $\hat{\boldsymbol{\mathcal{V}}}$ (equivalently $\hat{\Gamma}_\beta$ as well since $\boldsymbol{\mathcal{V}} := \text{vec}\left(\Gamma_\beta^\top\right)$) can be numerically solved using the efficient projection procedure suggested in Liu, Ji, and Ye (2009). Finally with estimated $\hat{\Gamma}_\beta$, fund-level indices constructed as the exposure to underlying firm-level characteristics are deemed as mattering for common exposure to factor structure if the the Euclidean norm of corresponding row of $\hat{\Gamma}_\beta$ larger then 0. The corresponding results with factor structure specified as ranging from 2 to 6 are summarized as following

[Place Table 2 about here]

As we can see from Table 2 that Euclidean norm of factor loadings on "value" (`BM_w`) and "size" (`Size_w`) at fund-level are all significantly positive in terms of magnitude, This suggests that based on the extended ℓ_1/ℓ_q-regularized IPCA, among the currently investigated 16 characteristics (probably the most prominent ones among all the 208 documented firm-level characteristics), if there indeed exists exposure of fund investment to common factors, conventional theory as in Fama and French (1992, 1993) survives. In comparison to the previous analysis in determining which characteristics of assets hold by fund matters for the exposure of fund investment based on established statistical testing framework in IPCA, 12-month as the characteristic measuring the differences of assets hold by funds survives in the sense that fund managers may still pay attention to momentum-related strategies in practical investment. Moreover, this also manifests that our proposed ℓ_1/ℓ_q-regularized IPCA may serve as one referenced alternative as the supplement, which suggest a broader view.

Remark 4.1 *It is well known that (see Park and Casella, 2008) regularized least squares optimization is closely connected with posterior analysis from Bayesian perspective. The suggested optima from (24) can be compared in analogy to optima implied from posterior mode by regarding imposed restrictions as the corresponding prior information specified from Bayesian perspective. In this*

regard, tuning parameter in (24) such as λ reflects the degree of informativeness associated with specified prior information and the smaller λ is, the less significantly the prior information specified matters.

Essentially as the tuning parameter, we follow the suggested routine in Liu, Ji, and Ye (2009) such that we calculate λ_{max} as the maximal value above which the objective function should obtain zero solution. With this automatically estimated λ_{max}, the optimization problem is regularized via $\lambda \times \lambda_{max}$.

4.4 Exposure to observable factors or not

In this section we report Wald-type statistic W_δ and the corresponding bootstrapped p-value for checking whether loading characterized by Γ_δ which maps some observable factors to fund-level characteristic space is significantly way from $\mathbf{0}$. We empirically find that when the number of latent factors is specified as $K = 3$ and Fama-French 3-factors are nested with latent factors, equity funds considered in our investigated sample are not significantly exposed to benchmark Fama-French 3-factor structure, in other word, we do not reject the null hypothesis that $\Gamma_\delta = \mathbf{0}$ ($W_\delta = 1.1345$ and p-value $= 0.8030$). Besides, we have also checked the exposure to Fama-French 3-factor structure plus momentum factor and we still do not reject the null-hypothesis $\Gamma_\delta = \mathbf{0}$ ($W_\delta = 1.4446$ and p-value $= 0.7360$). Besides, we also investigate the exposure to q-factors and expected growth factor proposed in Hou, Xue, and Zhang (2015) [10] but we still cannot reject the null-hypothesis ($W_\delta = 3.1779$ and p-value $= 0.7220$).

4.5 IPCA Performance measures

Two different easy-to-implement measures are exploited for measuring IPCA performance in this fund setting. The first one refers to "Total R^2": the fraction of variance in fund returns described by exposure to common factors.

$$\text{Total } R^2 = 1 - \frac{\sum_{i,t}\left(r_{i,t+1} - \hat{\beta}_{i,t}\hat{f}_{t+1}\right)^2}{\sum_{i,t} r_{i,t+1}^2} = 1 - \frac{\sum_{i,t}\left(r_{i,t+1} - z_{i,t}^\top\left(\hat{\Gamma}_\alpha + \hat{\Gamma}_\beta \hat{f}_{t+1}\right)\right)^2}{\sum_{i,t} r_{i,t+1}^2} \tag{25}$$

The second one refers to "predictive R^2": the fraction of variance in fund return described by conditional expected returns coming from exposure to common factors

$$\text{Predictive } R^2 = 1 - \frac{\sum_{i,t}\left(r_{i,t+1} - \hat{\beta}_{i,t}\hat{\lambda}\right)^2}{\sum_{i,t} r_{i,t+1}^2} = 1 - \frac{\sum_{i,t}\left(r_{i,t+1} - z_{i,t}^\top\left(\hat{\Gamma}_\alpha + \hat{\Gamma}_\beta \hat{\lambda}\right)\right)^2}{\sum_{i,t} r_{i,t+1}^2} \tag{26}$$

where $\hat{\gamma}$ refers to the vector of estimated risk prices, thus the time-series mean of estimated factors, $\hat{\gamma} = \frac{1}{T}\sum_t \hat{f}_t$. We summarize the corresponding results in the following table.

[10] We are grateful to Professor Lu Zhang for maintaining and releasing their data at http://global-q.org/factors.html.

[Place Table 3 about here]

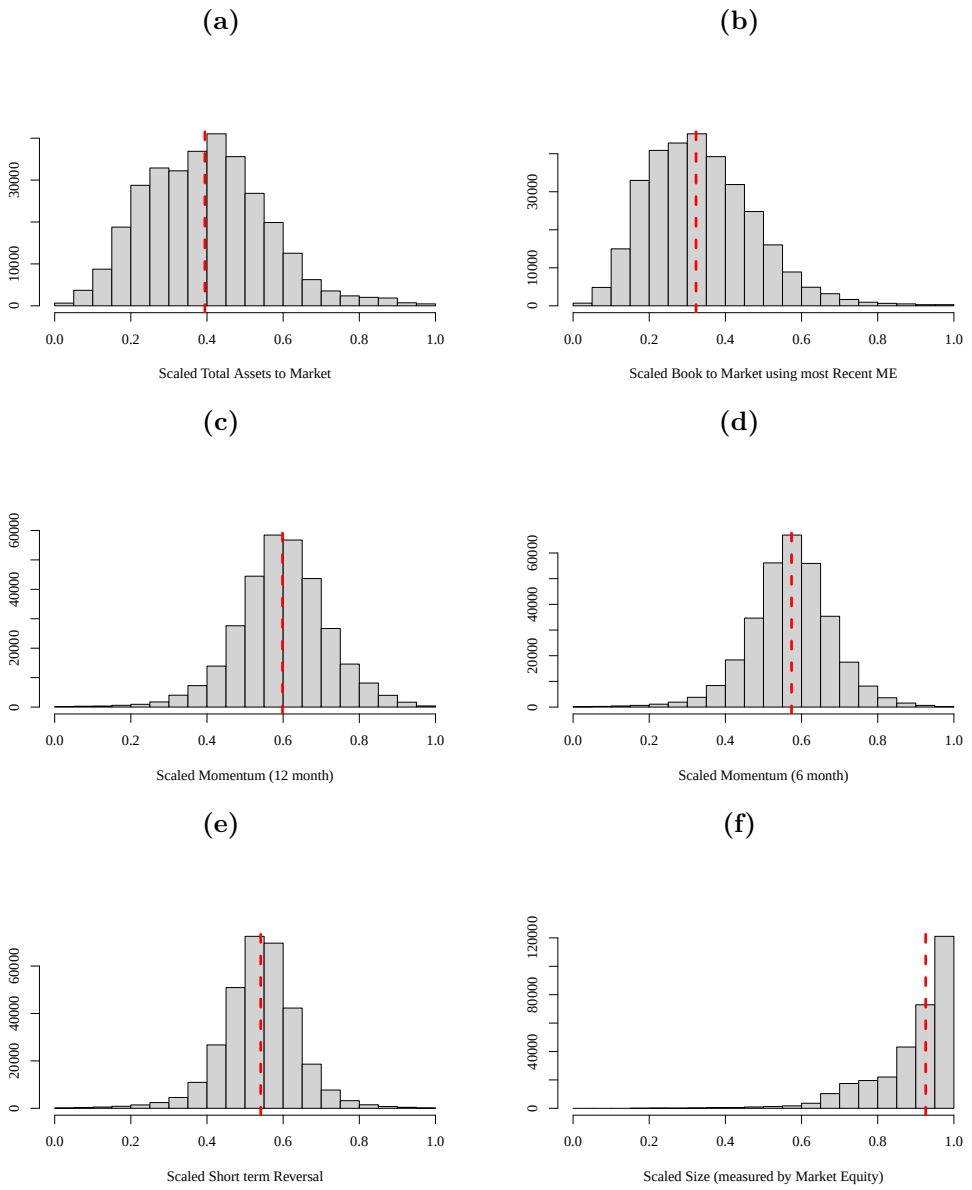
In the above table, we summarizes total R^2s and predictive R^2s corresponding different model specifications identified by number of factors and specification indicating whether to encompass constant intercept column in IPCA factor loading. For all the different cases with factor number specified ranging from 2 to 6, we do not reject null hypothesis $\Gamma_\alpha = \mathbf{0}$ based on the bootstrapped p-value associated with W_α statistic. This finding basically suggests that in modelling funds' exposure to cross-sectional information of holding assets using IPCA for accommodating dynamic factor loading, it is relatively not that serious by retaining focus on restricted model with $\Gamma_\alpha = \mathbf{0}$.

5 Conclusion

In this paper, we follow the standard routine in extant literature constructing a set of fund-level indices measuring the exposure of fund to the characteristics of assets hold by each fund and later empirically examine how funds are exposed to those firm-level characteristics (anomalies) based our constructed fund-level indices. Based on our constructed fund-level indices, our empirical analysis using standard IPCA (Instrumented Principal Component Analysis) and our extended ℓ_1/ℓ_q-regularized IPCA suggests that investment associated with equity funds are relatively more exposed at the "Value"-related and "Size"-related level if there exists certain exposure of fund to common factor structure. To sum up, the fund-level index constructed in this paper may suggest one alternative way to measure how equity fund is different from each other in terms of the exposure to characteristics of assets hold by funds. Besides, the empirical work that we conduct in this paper also extends the practical application of IPCA along with our proposed ℓ_1/ℓ_q-regularized IPCA by incorporating shrinkage using regularized norm rather than global optimization associated with ALS (alternating least squares) methodology. This extended IPCA may serve as one alternative supplement of the standard IPCA methodology.

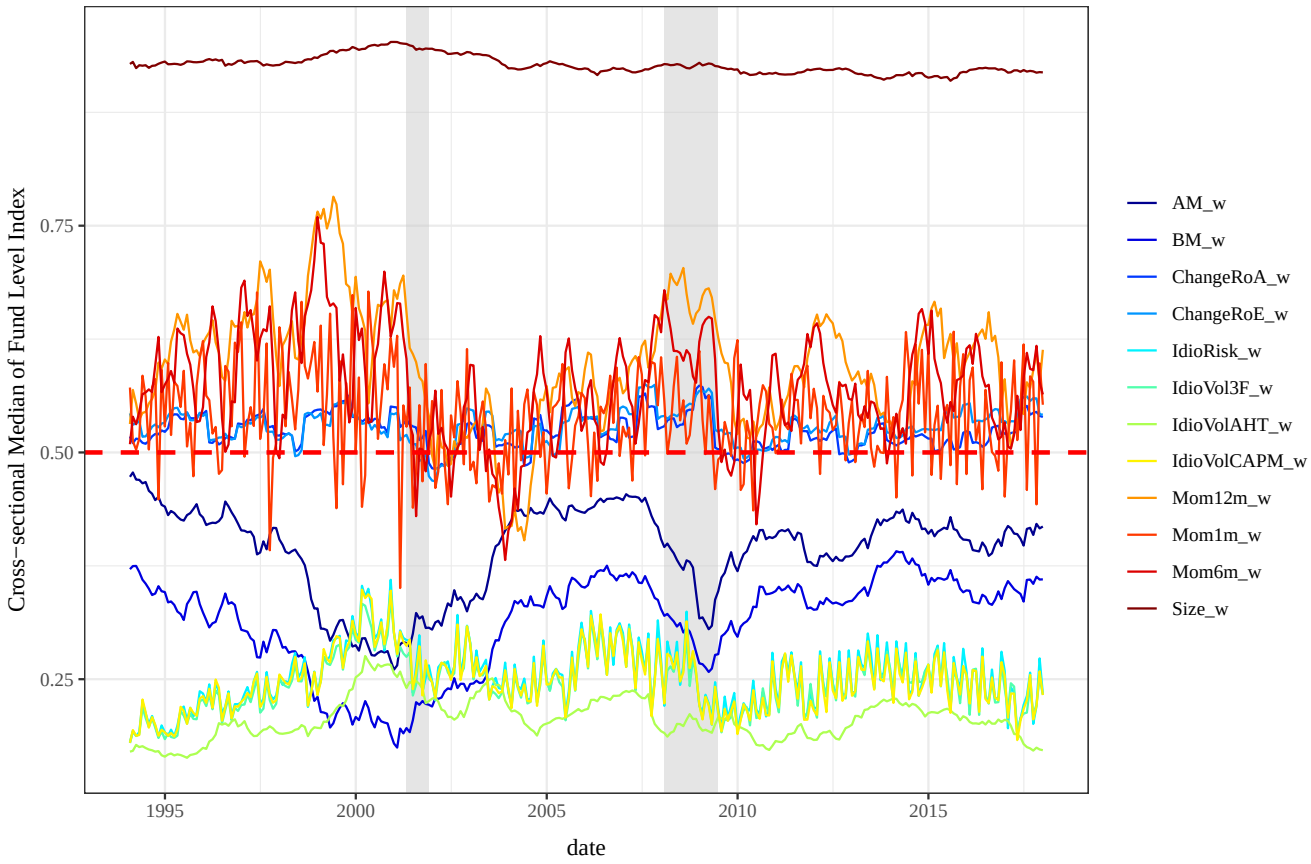
Figures and Tables

Figure 1



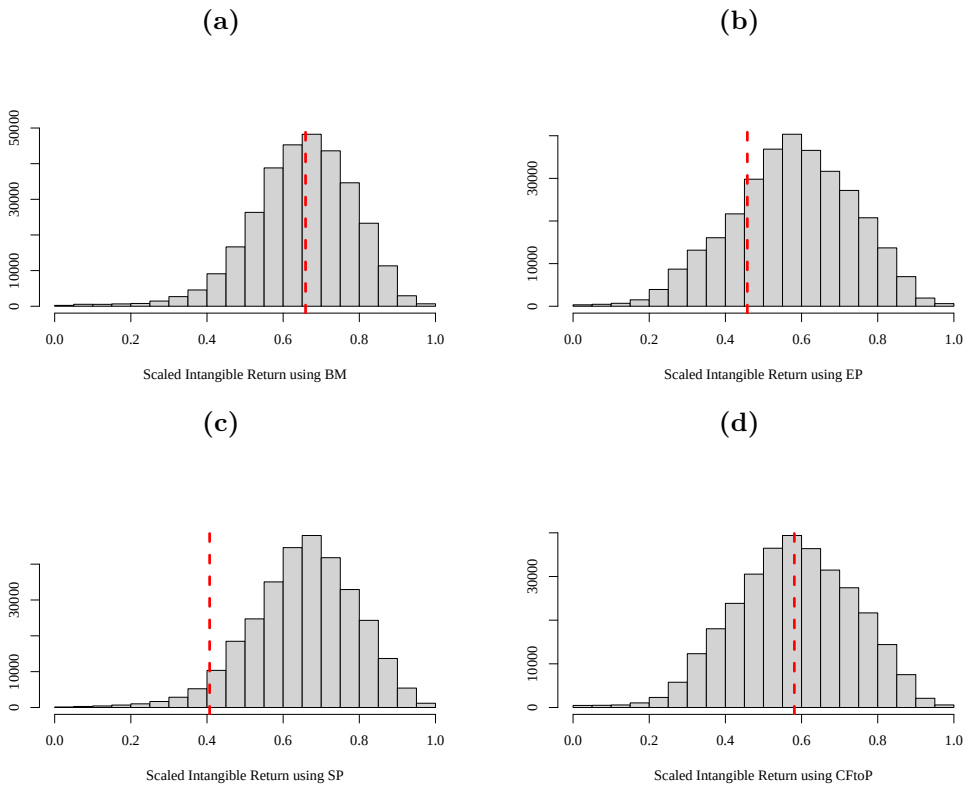
Note: Each histogram collected in panels indexed from (a) to (f) refers to the fund distribution on specific fund-level index constructed from corresponding firm-level characteristic, which is essentially the weighted average of specific firm-level characteristic as we have discussed in the main context. What demonstrated here correspond to 6 standard firm-level characteristics: AM (Assets to market equity); BM (Book to market equity using most recent ME); Mom12m (12-month momentum); Mom6m (6-month momentum); Mom1m (Short term reversal); Size (Measured by market equity).

Figure 2



Note: In the above figure, we demonstrate time-varying cross-sectional median associated with several major anomalies to which our constructed fund-level indices are exposed. The gray shaded area refers to NBER recession period over the selected sample from January 1994 to December 2017. We identify each anomaly at fund level as our constructed fund-level index with suffix `_w` attached to each anomaly names. The main implications from this figure are summarized as following: (1) For fund managers in the U.S. equity market, they persistently intend to hold large assets (thus high `Size_w`); (2) Assets hold by fund managers are persistently those assets exposed relatively less to idiosyncratic volatilities.

Figure 3

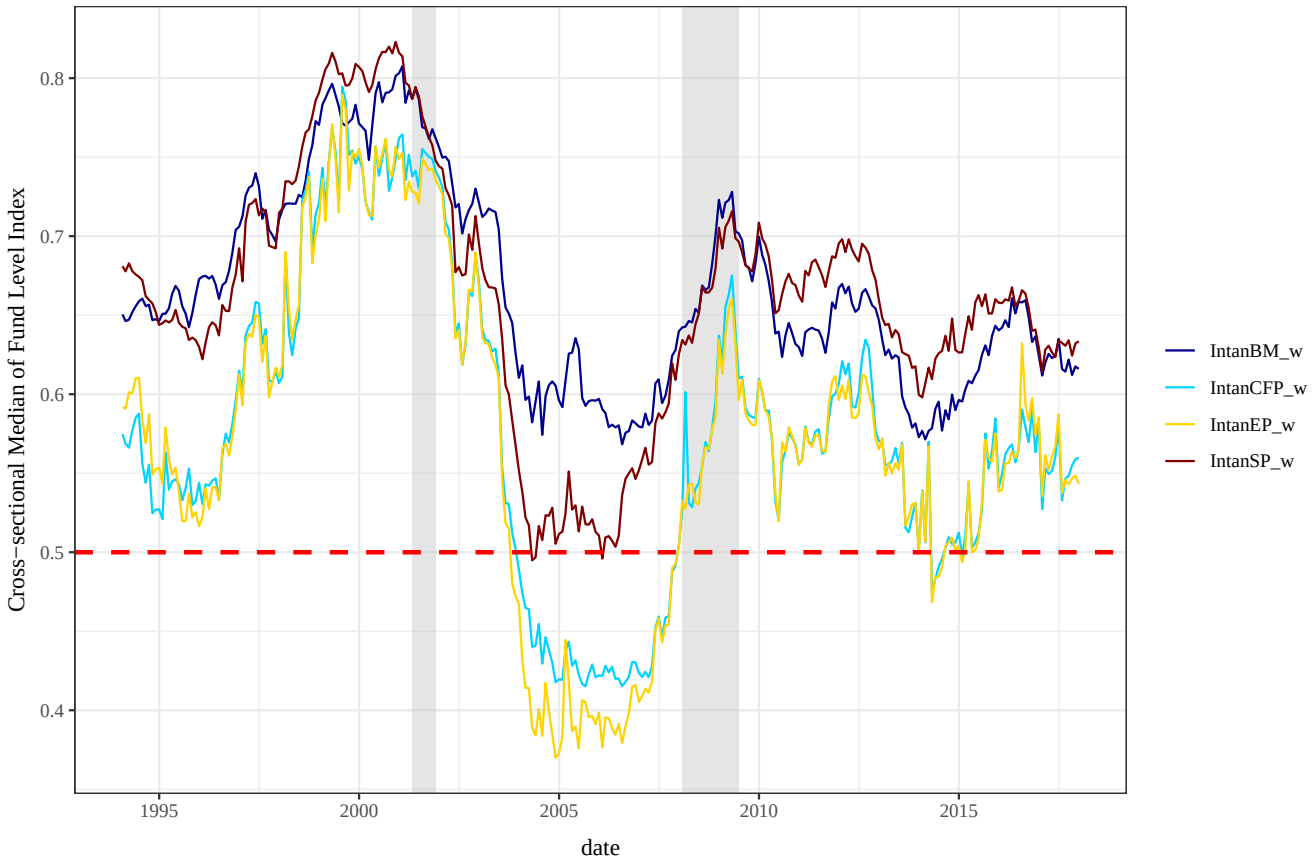


Note: DT2006 demonstrates that the following equation holds in general,

$$x_t = x_{t-\tau} + r^x(t-\tau, t) - r(t-\tau, t)$$

where x denotes log fundamental-price ratio (thus (a) log book-to-market ratio; (b) log equity-price ratio; (c) log sales-to-price ratio; (d) log cash flow-to-price ratio) while $r^x(t-\tau, t)$ refers to the log fundamental-price ratio return over the past τ periods and $r(t-\tau, t)$ refers to the log stock return over the past τ periods respectively as defined in DT2006. It is possible to roughly regard the tangible returns associated with x as the fitted component of cross-sectional regression induced from this equation while the intangible returns associated with x refers to the corresponding regression residuals. For more detailed and comprehensive discussions, please refer to the original paper of DT2006. Accordingly, if we treat the these intangible returns as characteristics at firm-level, it is naturally to construct the corresponding fund-level indices as we have discussed in the main context and the related distributions of funds on these constructed fund-level indices are demonstrated in each of the above panel in order. Red vertical line indicates the median.

Figure 4



Note: In the above figure, we specifically demonstrate time-varying cross-sectional median of anomalies associated with intangible returns as discussed in DT2006, to which our constructed fund-level indices are exposed. The gray shaded area refers to NBER recession period over the selected sample from January 1994 to December 2017. We identify each anomaly at fund level as our constructed fund level index with suffix `_w` attached to each anomaly names.

Table 1. Different Factor Structure Specification

	$K = 2$		$K = 3$		$K = 4$		$K = 5$		$K = 6$	
	$W_{\beta,l}$	p-value	$W_{\beta,l}$	p-value	$W_{\beta,l}$	p-value	$W_{\beta,l}$	p-value	$W_{\beta,l}$	p-value
AM_w	0.0027	0.5360	0.0681	0.0250**	0.3498	0.0000***	0.2793	0.0020***	0.5411	0.0000***
BM_w	0.2207	0.0080***	0.3623	0.0000***	0.4065	0.0000***	0.3725	0.0040***	0.3870	0.0080***
Mom12m_w	0.0052	0.4040	0.0109	0.6140	0.1670	0.1160	0.2010	0.1360	0.2299	0.1360
Mom6m_w	0.0167	0.2720	0.0331	0.3360	0.0748	0.2710	0.2722	0.0570*	0.2608	0.1190
Mom1m_w	0.0101	0.1700	0.0136	0.4190	0.0627	0.1750	0.0761	0.3100	0.0692	0.4360
Size_w	0.0128	0.1160	0.0324	0.1610	0.0400	0.1030	0.2389	0.0020***	0.2583	0.0000***
ChangeRoA_w	0.0083	0.5170	0.0111	0.7660	0.2195	0.1840	0.4331	0.0180**	0.6161	0.0020***
ChangeRoE_w	0.0091	0.4320	0.0181	0.5840	0.0404	0.5200	0.0672	0.5570	0.2239	0.1210
IdioRisk_w	0.1883	0.3260	0.2460	0.4430	0.2479	0.6940	0.2397	0.8200	0.3460	0.8460
IdioVol3F_w	0.6584	0.0000***	0.6576	0.0510*	0.6673	0.1520	0.6453	0.3070	0.6199	0.2910
IdioVolAHT_w	0.1027	0.0230**	0.3312	0.0020***	0.2920	0.0540*	0.3578	0.0100**	0.5345	0.0020***
IdioVolCAPM_w	0.3575	0.0440**	0.6365	0.2320	0.6902	0.2670	0.7021	0.4130	0.7372	0.4390
IntanBM_w	0.1680	0.0020***	0.1863	0.0140**	0.2349	0.0210**	0.2716	0.0480**	0.2986	0.0630*
IntanCFP_w	0.0706	0.2200	0.1205	0.2950	0.2060	0.3220	0.3006	0.2780	0.3014	0.4280
IntanEP_w	0.0806	0.0520*	0.1540	0.2670	0.1863	0.4490	0.3786	0.4490	0.4007	0.6340
IntanSP_w	0.0882	0.3720	0.1183	0.4280	0.1147	0.6490	0.1641	0.7200	0.1753	0.6130

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 2

	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 6$
AM_w	**0.0020**	0.0000	0.0000	0.0000	0.0000
BM_w	**0.0097**	**0.9999**	**0.9999**	**0.9996**	**0.9992**
Mom12m_w	0.0000	**0.9999**	**1.0000**	**1.0000**	**0.9999**
Mom6m_w	0.0000	0.0000	0.0000	0.0000	**1.0000**
Mom1m_w	0.0000	0.0000	0.0007	**0.9999**	**0.9999**
Size_w	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
ChangeRoA_w	0.0000	0.0000	0.0000	0.0000	0.0000
ChangeRoE_w	0.0000	0.0000	0.0000	0.0000	0.0000
IdioRisk_w	0.0000	0.0000	0.0000	0.0000	0.0000
IdioVol3F_w	0.0000	0.0000	0.0000	0.0000	0.0000
IdioVolAHT_w	0.0000	0.0000	0.0000	0.0000	0.0000
IdioVolCAPM_w	0.0000	0.0000	0.0000	0.0000	0.0000
IntanBM_w	0.0000	**0.0223**	**0.0142**	**0.0339**	**0.0448**
IntanCFP_w	0.0000	0.0000	0.0000	0.0000	0.0000
IntanEP_w	0.0000	0.0000	**1.0000**	**1.0000**	**1.0000**
IntanSP_w	**1.0000**	0.0000	0.0000	0.0000	**0.0082**
Pred.R^2	1.3563	1.3373	1.3315	1.3179	1.3168

Note: In this table, we report Euclidean norm of each row of estimated Γ_β from ℓ_1/ℓ_q-regularized IPCA for different factor structure specification such that the number of factor ranges from 2 to 6 and each column above refers to one specification. The last row refers to predictive R^2 to be discussed in subsection 4.5. Basically for each column, those entry with positive number implies the corresponding characteristics of assets hold by fund matter for the exposure of fund investment to the specified factor structure. These numbers are emphasizes in bold.

Table 3

		K				
		2	3	4	5	6
Individual Fund Testing						
Total R^2	$\Gamma_\alpha = \mathbf{0}$	66.50	66.73	66.94	67.13	67.30
	$\Gamma_\alpha \neq \mathbf{0}$	66.51	66.74	66.95	67.14	67.31
Pred. R^2	$\Gamma_\alpha = \mathbf{0}$	1.42	1.42	1.42	1.39	1.39
	$\Gamma_\alpha \neq \mathbf{0}$	1.39	1.39	1.39	1.35	1.37
Managed Portfolio Testing						
Total R^2	$\Gamma_\alpha = \mathbf{0}$	99.99	99.99	100.00	100.00	100.00
	$\Gamma_\alpha \neq \mathbf{0}$	99.99	99.99	100.00	100.00	100.00
Pred. R^2	$\Gamma_\alpha = \mathbf{0}$	2.57	2.56	2.58	2.53	2.54
	$\Gamma_\alpha \neq \mathbf{0}$	2.51	2.51	2.51	2.46	2.49
Intercept Γ_α Testing						
W_α p-value		0.84	0.76	0.83	0.79	0.61

Note: In this table, we summarize the IPCA performance in terms of total R^2s and predictive R^2s. Besides, we report p-values associated with Wald statistic W_α for testing statistical significance of intercept term contained in IPCA factor loading. All the numbers as the entries of above table except for the last row reporting p-value are corresponding R^2s in percentage.

References

Ali, A., X. Chen, T. Yao, and T. Yu (2008): "Do Mutual Funds Profit from the Accruals Anomaly?," *Journal of Accounting Research*, 46(1), 1–26. [Cited on page 10.]

Bai, J. (2003): "Inferential Theory for Factor Models of Large Dimensions," *Econometrica*, 71(1), 135–171. [Cited on page 2.]

Bai, J., and S. Ng (2002): "Determing the Number of Factors in Approximate Factor Models," *Econometrica*, 70(1), 191–221. [Cited on page 2.]

——— (2013): "Principal components estimation and identification of static factors," *Journal of Econometrics*, 176(1), 18–29. [Cited on page 2.]

Barber, B. M., X. Huang, and T. Odean (2016): "Which Factors Matter to Investors? Evidence from Mutual Fund Flows," *The Review of Financial Studies*, 29(10), 2600–2642. [Cited on page 11.]

Büchner, M. (2021): "What Drives Asset Holdings ? Commanality in Investor Demand," working paper. [Cited on page 2.]

Chamberlain, G., and M. Rothschild (1983): "Arbitrage, Factor Structure, and Mean-Variance Analysis on Large Asset Markets," *Econometrica*, 51(5), 1281–1304. [Cited on page 1.]

Chen, A. Y., and T. Zimmermann (2020): "Open Source Cross-Sectional Asset Pricing," Working paper, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3604626. [Cited on pages 1, 2, 9, and 12.]

Chen, Y. (2019): "Estimating Expected Return Function Nonparametrically: Based on BART," Working paper. [Cited on pages 1 and 9.]

Connor, G., and R. A. Korajczyk (1986): "Performance measurement with the arbitrage pricing theory: A new framework for analysis," *Journal of Financial Economics*, 15(3), 373–394. [Cited on page 1.]

Daniel, K., and S. Titman (2006): "Market Reactions to Tangible and Intangible Information," *The Journal of Finance*, 61(4), 1605–1643. [Cited on pages 12, 20, and 21.]

Demiguel, V., A. Martín, F. J. Nogales, and R. Uppal (2020): "A Transaction-Cost Perspective on the Multitude of Firm Characteristics," *The Review of Financial Studies*, 33(5), 2180–2122. [Cited on pages 1 and 9.]

Eisfeldt, A. L., E. Kim, and D. Papanikolaou (2020): "Intangible Value," Working paper. [Cited on page 12.]

Fama, E. F., and K. R. French (1992): "The Cross-Section of Expected Stock Returns," *The Journal of Finance*, 47(2), 427–465. [Cited on pages 1, 3, 9, and 15.]

——— (1993): "Common Risk Factors in the Returns on Stocks and Bonds," *Journal of Financial Economics*, 33(1), 3–56. [Cited on pages 1, 3, 9, and 15.]

——— (1996): "Multifactor Explanations of Asset Pricing Anomalies," *The Journal of Finance*, 51(1), 55–84. [Cited on pages 1 and 9.]

——— (1997): "Industry costs of equity," *Journal of Financial Economics*, 43(2), 153–193. [Cited on page 9.]

——— (2010): "Luck versus Skill in the Cross-Section of Mutual Fund Returns," *The Journal of Finance*, 65(5), 1915–1947. [Cited on page 1.]

——— (2015): "A Five-factor Asset Pricing Model," *Journal of Financial Economics*, 116(1), 1 – 22. [Cited on page 1.]

Fan, J., Y. Liao, and W. Wang (2016): "Projected Principal Component Analysis in Factor Models," *The Annals of Statistics*, 44(1), 219–254. [Cited on page 2.]

Freybergerk, J., A. Neuhierl, and M. Weber (2019): "Dissecting Characteristics Nonparametrically," Working paper, forthcoming in *The Review of Financial Studies*. [Cited on pages 1 and 9.]

Geweke, J. (1977): "The Dynamic Factor Analysis of Economic Time Series," Latent variables in socio-economic models. [Cited on page 2.]

Green, J., J. R. M. Hand, and X. F. Zhang (2017): "The Characteristics that Provide Independent Information about Average U.S. Monthly Stock Returns," *The Review of Financial Studies*, 30(12), 4389–4436. [Cited on pages 1 and 9.]

Gu, S., B. Kelly, and D. Xiu (2019): "Empirical Asset Pricing via Maching Learning," Working paper, forthcoming in the *The Review of Financial Studies*. [Cited on pages 1, 3, and 9.]

Haddad, V., S. Kozak, and S. Santosh (2020): "Factor Timing," *Review of Financial Studies*, 33(5), 1980–2018. [Cited on page 1.]

Harvey, C. R., and Y. Liu (2014): "Evaluationg Trading Strategies," *Journal of Portfolio Management*, 40(5), 108–118. [Cited on pages 1 and 9.]

——— (2015): "Backtesting," *Journal of Portfolio Management*, 42(1), 13–28. [Cited on pages 1 and 9.]

——— (2020): "Luck versus Skill in the Cross-Section of Mutual Fund Returns: Reexamining the Evidence," Working paper. [Cited on page 1.]

Harvey, C. R., Y. Liu, and H. Zhu (2016): "...and the Cross-Section of Expected Returns," *The Review of Financial Studies*, 29(1), 5–68. [Cited on pages 1 and 9.]

Hoberg, G., N. Kumar, and N. Prabhala (2017): "Mutual Fund Competition, Managerial Skill, and Alpha Persistence," *The Review of Financial Studies*, 31(5), 1896–1929. [Cited on pages 2 and 8.]

Hou, K., C. Xue, and L. Zhang (2015): "Digesting Anomalies: An Investment Approach," *The Review of Financial Studies*, 28(3), 650–705. [Cited on pages 1 and 16.]

——— (2018): "Replicating Anomalies," *The Review of Financial Studies*, 33(5), 2019–2133. [Cited on page 1.]

Kacperczyk, M., C. Sialm, and L. Zheng (2005): "On the Industry Concentration of Actively Managed Equity Mutual Funds," *The Journal of Finance*, 60(4), 1983–2011. [Cited on page 8.]

Kacperczyk, M., C. Sialm, and L. Zheng (2006): "Unobserved Actions of Mutual Funds," *The Review of Financial Studies*, 21(6), 2379–2416. [Cited on pages 2 and 8.]

Kelly, B. T., T. J. Moskowitz, and S. Pruitt (2021): "Understanding Momentum and Reversal," *Journal of Financial Economics*. [Cited on page 2.]

Kelly, B. T., D. Palhares, and S. Pruitt (2020): "Modeling Corporate Bond Returns," Working paper. [Cited on pages 2 and 3.]

Kelly, B. T., S. Pruitt, and Y. Su (2019): "Characteristics are covariances: A unified model of risk and return," *Journal of Financial Economics*, 134(3), 501–524. [Cited on pages 1, 2, 3, 4, 5, and 6.]

Kelly, B. T., S. Pruitt, and Y. Su (2020): "Instrumented Principal Component Analysis," Working paper. [Cited on pages 1, 2, 3, 4, and 5.]

Kosowski, R., A. Timmermann, R. Wermers, and H. White (2006): "Can Mutual Fund "Stars" Really Pick Stocks? New Evidence from a Bootstrap Analysis," *The Journal of Finance*, 61(6), 2551–2595. [Cited on page 1.]

Kozak, S. (2020): "Kernel Trick for the Cross Section," Working paper. [Cited on pages 1, 2, 9, and 13.]

Kozak, S., S. Nagel, and S. Santosh (2018): "Interpreting Factor Models," *The Journal of Finance*, 73(3), 1183–1223. [Cited on page 1.]

——— (2020): "Shrinking the Cross Section," *Journal of Financial Economics*, 135(2), 271–292. [Cited on pages 1, 2, and 9.]

Lettau, M., S. C. Ludvigson, and P. Manoel (2021): "Characteristics of Mutual Fund Portfolios: Where are the Value Funds," NBER Working paper No. w25381. [Cited on page 11.]

Lettu, M., and M. Pelger (2020a): "Estimating Latent Asset-Pricing Factors," forthcoming in *The Journal of Econometrics*. [Cited on page 1.]

——— (2020b): "Factors that Fit the Time-Series and Cross-Section of Stock Returns," *Review of Financial Studies*, 33(5), 2274–2325. [Cited on page 1.]

Li, B., and A. Rossi (2021): "Selectiong Mutual Funds from the Stokcs They Hold: A Machine Learning Approach," working paper. [Cited on page 2.]

Liu, J., S. Ji, and J. Ye (2009): "SLEP: Sparse Learning with Efficient Projections," Technical Report. [Cited on pages 15 and 16.]

Magnus, J. R., and H. Neudecker (2007): *Matrix Differential Calculations: with Applications in Statistics and Econometrics,* 3rd edition, Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, NJ. [Cited on page A-1.]

Park, T., and G. Casella (2008): "The Bayesian Lasso," *Journal of the American Statistical Association*, 103(482), 681–686. [Cited on page 15.]

Sargent, T., and C. Sims (1977): "Business cycle modeling without pretending to have too much a priori economic theory," Working Papers 55, Federal Reserve Bank of Minneapolis. [Cited on page 2.]

Stock, J. H., and M. W. Watson (2002): "Forecasting Using Principal Components From a Large Number of Predictors," *Journal of the American Statistical Association*, 97(460), 1167–1179. [Cited on page 2.]

Welch, I. (2019): "Reproducing, Extending, Updating, Replicationg, Reexamining, and Reconciling," *Critical Finance Review*, 8(1-2), 301–304. [Cited on page 9.]

Appendix

A Auxiliary Proofs

Next we proceed to check the first order derivative taken with respect to $\text{vec}\left(\Gamma_\beta^\top\right)$ with fixed $\hat{f}_{t+1}$. The associated F.O.C. for each t writes as following. Moreover, to facilitate our later discussion we carry over our discussion by rewriting the formula as following

$$\mathbf{r}_{t+1} - \mathbf{Z}_t\Gamma_\beta f_{t+1} = \mathbf{r}_{t+1} - \text{vec}\left(f_{t+1}^\top\Gamma_\beta^\top\mathbf{Z}_t^\top\right) = \mathbf{r}_{t+1} - \left(\mathbf{Z}_t \otimes f_{t+1}^\top\right)\text{vec}\left(\Gamma_\beta^\top\right)$$

which implies that the F.O.C. taken with respect to $\Gamma_\beta^\top$ can be equivalently represented as following as well [1]

$$\begin{aligned}
\sum_{t=1}^{T-1}\left(\mathbf{Z}_t \otimes f_{t+1}^\top\right)^\top\mathbf{r}_{t+1} &= \sum_{t=1}^{T-1}\left[\left(\mathbf{Z}_t \otimes f_{t+1}^\top\right)^\top\left(\mathbf{Z}_t \otimes f_{t+1}^\top\right)\right]\text{vec}\left(\Gamma_\beta^\top\right) \\
&= \sum_{t=1}^{T-1}\left[\left(\mathbf{Z}_t^\top \otimes f_{t+1}\right)\left(\mathbf{Z}_t \otimes f_{t+1}^\top\right)\right]\text{vec}\left(\Gamma_\beta^\top\right) \\
&= \sum_{t=1}^{T-1}\left(\mathbf{Z}_t^\top\mathbf{Z}_t \otimes f_{t+1}f_{t+1}^\top\right)\text{vec}\left(\Gamma_\beta^\top\right)
\end{aligned}$$

Thus

$$\begin{aligned}
\text{vec}\left(\hat{\Gamma}_\beta^\top\right) &= \left(\sum_{t=1}^{T-1}\mathbf{Z}_t^\top\mathbf{Z}_t \otimes f_{t+1}f_{t+1}^\top\right)^{-1}\left(\sum_{t=1}^{T-1}\left(\mathbf{Z}_t^\top \otimes f_{t+1}\right)\mathbf{r}_{t+1}\right) \\
&= \left(\sum_{t=1}^{T-1}\mathbf{Z}_t^\top\mathbf{Z}_t \otimes f_{t+1}f_{t+1}^\top\right)^{-1}\left(\sum_{t=1}^{T-1}\left(\mathbf{Z}_t^\top \otimes r_{t+1}\right)f_{t+1}\right)
\end{aligned}$$

[1] Keep in mind the following property covered in equation (4) in chapter 2 of Magnus and Neudecker (2007),

$$(A \otimes B)(C \otimes D) = AC \otimes BD \tag{F.1}$$

if AC and BD exists.

B 16 Basic Anomalies plus 3 Basic Fund Characteristics

Table B.1. Different Factor Structure Specification

	K = 2		K = 3		K = 4		K = 5		K = 6	
	$W_{\beta,l}$	p-value	$W_{\beta,l}$	p-value	$W_{\beta,l}$	p-value	$W_{\beta,l}$	p-value	$W_{\beta,l}$	p-value
AM_w	0.0126	0.2560	0.0816	0.0400**	0.1938	0.0410**	0.2857	0.0260**	0.5488	0.0000***
BM_w	0.1422	0.0080***	0.1802	0.0190**	0.3800	0.0030***	0.3690	0.0210**	0.4077	0.0210**
Mom12m_w	0.0091	0.2360	0.0277	0.2440	0.0355	0.3250	0.0652	0.3360	0.0357	0.7150
Mom6m_w	0.0332	0.0630*	0.0848	0.1000	0.1155	0.0950*	0.1094	0.1690	0.0985	0.3270
Mom1m_w	0.0181	0.0240**	0.0369	0.0620*	0.0446	0.1260	0.0518	0.1970	0.0616	0.3220
Size_w	0.1710	0.0000***	0.2435	0.0000***	0.2207	0.0020***	0.3227	0.0000***	0.2495	0.0050***
ChangeRoA_w	0.0082	0.3750	0.0166	0.5250	0.0351	0.5800	0.0928	0.2550	0.1534	0.3970
ChangeRoE_w	0.0166	0.1720	0.0476	0.1770	0.0521	0.3390	0.0631	0.4730	0.0723	0.6490
IdioRisk_w	0.2206	0.2810	0.1796	0.6380	0.3459	0.5410	0.3464	0.7130	0.5715	0.6620
IdioVol3F_w	0.0173	0.7950	0.4423	0.0840*	0.6047	0.0720*	0.6202	0.0820*	0.7138	0.1490
IdioVolAHT_w	0.0636	0.0340**	0.0580	0.2210	0.2252	0.0460**	0.3019	0.0950*	0.4404	0.0060***
IdioVolCAPM_w	0.3828	0.3030	0.4308	0.5120	0.4582	0.6830	0.6423	0.4090	0.7839	0.3260
IntanBM_w	0.1011	0.0090***	0.1610	0.0960*	0.1987	0.0190**	0.2429	0.0230**	0.3915	0.0230**
IntanCFP_w	0.2479	0.0540*	0.2971	0.1750	0.2904	0.1770	0.3278	0.2880	0.5141	0.2730
IntanEP_w	0.1125	0.1040	0.1517	0.2580	0.1766	0.3050	0.3759	0.2890	0.2853	0.5630
IntanSP_w	0.0670	0.0720*	0.0612	0.2640	0.1037	0.3970	0.1945	0.3400	0.0754	0.9000

Table B.1. Different Factor Structure Specification *(continued)*

	K = 2		K = 3		K = 4		K = 5		K = 6	
	$W_{\beta,l}$	p-value	$W_{\beta,l}$	p-value	$W_{\beta,l}$	p-value	$W_{\beta,l}$	p-value	$W_{\beta,l}$	p-value
mtna_normalized	0.0077	0.0110**	0.0334	0.0060***	0.0195	0.0010***	0.0325	0.0000***	0.0468	0.0000***
flow_normalized	0.2416	0.0000***	0.2361	0.0000***	0.2589	0.0000***	0.2883	0.0000***	0.2125	0.0000***
age_normalized	0.1269	0.0000***	0.2299	0.0000***	0.2411	0.0000***	0.2675	0.0000***	0.3372	0.0000***

C Comprehensive 40 characteristics

Table C.1. Different Factor Structure Specification 40

	K = 2		K = 3		K = 4		K = 5		K = 6	
	$W_{\beta,l}$	p-value	$W_{\beta,l}$	p-value	$W_{\beta,l}$	p-value	$W_{\beta,l}$	p-value	$W_{\beta,l}$	p-value
Size_w	0.0315	0.3290	0.0569	0.1430	0.1407	0.3460	0.2311	0.2120	0.2193	0.3710
BM_w	0.0398	0.5290	0.1736	0.0370**	0.1469	0.5460	0.0968	0.8190	0.0961	0.9060
GP_w	0.0406	0.2790	0.0859	0.4430	0.1957	0.0900*	0.0373	0.8530	0.0612	0.9450
Profitability_w	0.0458	0.2300	0.1221	0.0640*	0.0247	0.8090	0.0619	0.7500	0.1226	0.5880
PS_w	0.0008	0.6020	0.0006	0.6880	0.0099	0.2430	0.0052	0.6720	0.0148	0.3790
DebtIssuance_w	0.0242	0.2330	0.0417	0.3600	0.0950	0.0800*	0.1171	0.1540	0.1284	0.2520
ShareRepurchase_w	0.0013	0.8570	0.0098	0.7350	0.0338	0.4750	0.0284	0.6900	0.0861	0.4020
ShareIss1Y_w	0.0261	0.2290	0.0336	0.3020	0.0043	0.9650	0.0366	0.6830	0.0605	0.6780
Accruals_w	0.0069	0.3010	0.1319	0.0100**	0.0235	0.2300	0.0209	0.5250	0.0397	0.4880
AssetGrowth_w	0.0047	0.6430	0.0444	0.5470	0.0257	0.6050	0.0673	0.4010	0.0823	0.4340
ChAssetTurnover_w	0.0063	0.3090	0.1380	0.0020***	0.0120	0.4690	0.0220	0.4400	0.0171	0.7210
DivYield_w	0.0039	0.1150	0.0070	0.0530*	0.0066	0.3030	0.0040	0.6530	0.0084	0.4550
EP_w	0.0107	0.5570	0.0313	0.3710	0.0101	0.9390	0.0195	0.9250	0.0495	0.8420
CF_w	0.0801	0.1320	0.1144	0.1660	0.0721	0.6080	0.3260	0.0900*	0.3015	0.1620
NOA_w	0.0424	0.0860*	0.1050	0.0130**	0.1042	0.0950*	0.1246	0.1120	0.1428	0.1850
Investment_w	0.0065	0.3080	0.0203	0.4960	0.0089	0.6200	0.0024	0.9760	0.0088	0.8950

Table C.1. Different Factor Structure Specification 40 *(continued)*

	K = 2		K = 3		K = 4		K = 5		K = 6	
	$W_{\beta,l}$	p-value	$W_{\beta,l}$	p-value	$W_{\beta,l}$	p-value	$W_{\beta,l}$	p-value	$W_{\beta,l}$	p-value
InvGrowth_w	0.0057	0.3770	0.0241	0.2350	0.0035	0.8850	0.0133	0.6550	0.0067	0.9500
Leverage_w	0.1708	0.1980	0.3090	0.0080***	0.1497	0.5690	0.1816	0.6780	0.3142	0.6790
SP_w	0.1293	0.0470**	0.0281	0.5060	0.1458	0.1970	0.1672	0.2500	0.1796	0.4550
GrLTNOA_w	0.0243	0.1290	0.0929	0.0410**	0.0216	0.4850	0.0159	0.7450	0.0136	0.9140
Mom6m_w	0.0436	0.2600	0.0564	0.3430	0.0143	0.8870	0.1004	0.4700	0.0831	0.7310
IndMom_w	0.0044	0.6300	0.0178	0.4270	0.0185	0.6990	0.0271	0.7660	0.0142	0.9800
ShortInterest_w	0.0098	0.4120	0.0137	0.4690	0.1490	0.0300**	0.0638	0.4110	0.0891	0.4150
Mom12m_w	0.0080	0.7410	0.0642	0.2180	0.0837	0.4810	0.0827	0.7690	0.2310	0.2850
Mom1m_w	0.0066	0.4630	0.0084	0.6700	0.0030	0.9630	0.0155	0.8320	0.0135	0.9410
Mom18m13m_w	0.0041	0.6920	0.0111	0.7300	0.0137	0.7740	0.0761	0.3250	0.1216	0.2890
EarningsSurprise_w	0.0003	0.9430	0.0686	0.1650	0.0082	0.8360	0.0247	0.7180	0.0711	0.4070
ChangeRoA_w	0.0031	0.7810	0.0617	0.4480	0.0406	0.6440	0.0460	0.7140	0.1159	0.4450
ChangeRoE_w	0.0021	0.8020	0.0229	0.8090	0.0305	0.6250	0.0062	0.9850	0.0291	0.9320
IdioRisk_w	0.1279	0.7970	0.1786	0.8350	0.5263	0.8390	0.6479	0.6280	0.6724	0.8820
IdioVol3F_w	0.2821	0.4620	0.4718	0.1860	0.5138	0.6450	0.5106	0.8120	0.5147	0.8920
IdioVolAHT_w	0.0408	0.5820	0.0594	0.4550	0.1467	0.5910	0.3483	0.2880	0.3306	0.5180
IdioVolCAPM_w	0.3107	0.7820	0.2383	0.8200	0.5688	0.8050	0.6225	0.8600	0.7196	0.9620
CompEquIss_w	0.0167	0.2470	0.0083	0.5780	0.0167	0.6870	0.1099	0.1200	0.1047	0.3350

Table C.1. Different Factor Structure Specification 40 *(continued)*

	$K = 2$		$K = 3$		$K = 4$		$K = 5$		$K = 6$	
	$W_{\beta,l}$	p-value	$W_{\beta,l}$	p-value	$W_{\beta,l}$	p-value	$W_{\beta,l}$	p-value	$W_{\beta,l}$	p-value
CompositeDebtIssuance_w	0.0063	0.3030	0.0143	0.3990	0.0081	0.7180	0.0407	0.2380	0.0240	0.6220
ShareVol_w	0.0384	0.0940*	0.0233	0.4390	0.0881	0.0840*	0.0811	0.1820	0.0435	0.6460
EquityDuration_w	0.0105	0.7290	0.0023	0.9630	0.0915	0.4510	0.0779	0.6960	0.2372	0.3490
mtna_normalized	0.0112	0.0070***	0.0028	0.0070***	0.0416	0.0030***	0.0352	0.0090***	0.0339	0.0220**
flow_normalized	0.1730	0.0000***	0.0392	0.0040***	0.1710	0.0000***	0.1768	0.0000***	0.2386	0.0000***
age_normalized	0.1985	0.0000***	0.0661	0.0010***	0.2311	0.0000***	0.3276	0.0000***	0.3591	0.0000***

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

D Comprehensive List of Characteristics

Table D.1. Detailed Descriptions of used Anomalies

No.	Acronym	Authors	Year	Long Description	Journal
1	AbnormalAccruals	Xie	2001	Abnormal Accruals	AR
2	AbnormalAccrualsPercent	Hafzalla, Lundholm and Van Winkle	2011	Percent Abnormal Accruals	AR
3	Accruals	Sloan	1996	Accruals	AR
4	AccrualsBM	Bartov and Kim	2004	Book-to-market and accruals	RFQA
5	Activism1	Cremers and Nair	2005	Shareholder activism 1	JF
6	Activism2	Cremers and Nair	2005	Shareholder activism 2	JF
7	AdExp	Chan, Lakonishok and Sougiannis	2001	Advertising Expense	JF
8	AgeIPO	Ritter	1991	IPO and age	JF
9	AM	Fama and French	1992	Total assets to market	JF
10	AnalystValue	Frankel and Lee	1998	Analyst Value	JAE
11	AnnouncementReturn	Chan, Jegadeesh and Lakonishok	1996	Earnings announcement return	JF
12	AOP	Frankel and Lee	1998	Analyst Optimism	JAE
13	AssetGrowth	Cooper, Gulen and Schill	2008	Asset Growth	JF
14	Beta	Fama and MacBeth	1973	CAPM beta	JPE
15	BetaBDLeverage	Adrian, Etula and Muir	2014	Broker-Dealer Leverage Beta	JF
16	BetaFP	Frazzini and Pedersen	2014	Frazzini-Pedersen Beta	JFE
17	BetaLiquidityPS	Pastor and Stambaugh	2003	Pastor-Stambaugh liquidity beta	JPE
18	BetaTailRisk	Kelly and Jiang	2014	Tail risk beta	RFS
19	betaVIX	Ang et al.	2006	Systematic volatility	JF
20	BidAskSpread	Amihud and Mendelsohn	1986	Bid-ask spread	JFE

Table D.1. Detailed Descriptions of used Anomalies *(continued)*

No.	Acronym	Authors	Year	Long Description	Journal
21	BM	Rosenberg, Reid, and Lanstein	1985	Book to market using most recent ME	JF
22	BMdec	Rosenberg, Reid, and Lanstein	1985	Book to market using December ME	JPM
23	BookLeverage	Fama and French	1992	Book leverage (annual)	JF
24	BPEBM	Penman, Richardson and Tuna	2007	Leverage component of BM	JAR
25	BrandInvest	Belo, Lin and Vitorino	2014	Brand capital investment	RED
26	Cash	Palazzo	2012	Cash to assets	JFE
27	CashProd	Chandrashekar and Rao	2009	Cash Productivity	WP
28	CBOperProfNoLag	Ball et al.	2016	Cash-based operating profitability	JFE
29	CF	Lakonishok, Shleifer and Vishny	1994	Cash flow to market	JF
30	cfp	Desai, Rajgopal and Venkatachalam	2004	Operating Cash flows to price	AR
31	ChangeInRecommendation	Jegadeesh et al.	2004	Change in recommendation	JF
32	ChangeRoA	Hou, Xue and Zhang	2018	Change in Return on assets	RFS
33	ChangeRoE	Hou, Xue and Zhang	2018	Change in Return on equity	RFS
34	ChAssetTurnover	Soliman	2008	Change in Asset Turnover	AR
35	ChEQ	Lockwood and Prombutr	2010	Sustainable Growth	JFR
36	ChForecastAccrual	Barth and Hutton	2004	Change in Forecast and Accrual	RAS
37	ChInv	Thomas and Zhang	2002	Inventory Growth	RAS
38	ChInvIA	Abarbanell and Bushee	1998	Change in capital inv (ind adj)	AR
39	ChNAnalyst	Scherbina	2008	Decline in Analyst Coverage	ROF
40	ChNNCOA	Soliman	2008	Change in Net Noncurrent Operating Assets	AR
41	ChNWC	Soliman	2008	Change in Net Working Capital	AR
42	ChTax	Thomas and Zhang	2011	Change in Taxes	JAR

Table D.1. Detailed Descriptions of used Anomalies *(continued)*

No.	Acronym	Authors	Year	Long Description	Journal
43	CompEquIss	Daniel and Titman	2006	Composite equity issuance	JF
44	CompositeDebtIssuance	Lyandres, Sun and Zhang	2008	Composite debt issuance	RFS
45	ConsRecomm	Barber et al.	2002	Consensus Recommendation	JF
46	ConvDebt	Valta	2016	Convertible debt indicator	JFQA
47	Coskewness	Harvey and Siddique	2000	Coskewness	JF
48	CredRatDG	Dichev and Piotroski	2001	Credit Rating Downgrade	JF
49	CustomerMomentum	Cohen and Frazzini	2008	Customer momentum	JF
50	DebtIssuance	Spiess and Affleck-Graves	1999	Debt Issuance	JFE
51	DelBreadth	Chen, Hong and Stein	2002	Breadth of ownership	JFE
52	DelCOA	Richardson et al.	2005	Change in current operating assets	JAE
53	DelCOL	Richardson et al.	2005	Change in current operating liabilities	JAE
54	DelDRC	Prakash and Sinha	2012	Deferred Revenue	CAR
55	DelEqu	Richardson et al.	2005	Change in equity to assets	JAE
56	DelFINL	Richardson et al.	2005	Change in financial liabilities	JAE
57	DelLTI	Richardson et al.	2005	Change in long-term investment	JAE
58	DelNetFin	Richardson et al.	2005	Change in net financial assets	JAE
59	DivInd	Hartzmark and Salomon	2013	Dividends	JFE
60	DivInit	Michaely, Thaler and Womack	1995	Dividend Initiation	JF
61	DivOmit	Michaely, Thaler and Womack	1995	Dividend Omission	JF
62	DivYield	Naranjo, Nimalendran and Ryngaert	1998	Dividend Yield	JF
63	dNoa	Hirshleifer, Hou, Teoh, Zhang	2004	change in net operating assets	JAE
64	DolVol	Brennan, Chordia and Subrahmanyam	1998	Past trading volume	JFE

Table D.1. Detailed Descriptions of used Anomalies *(continued)*

No.	Acronym	Authors	Year	Long Description	Journal
65	DownForecast	Barber et al.	2002	Down forecast EPS	JF
66	EarnIncrease	Loh and Warachka	2012	Earnings streak indicator	MS
67	EarningsConsistency	Alwathainani	2009	Earnings growth for consistent growers	BAR
68	EarningsForecastDisparity	Da and Warachka	2011	Long vs short-term earnings expectations	JFE
69	EarningsSurprise	Foster, Olsen and Shevlin	1984	Earnings Surprise	AR
70	EarnSupBig	Hou	2007	Earnings surprise of big firms	RFS
71	EBM	Penman, Richardson and Tuna	2007	Enterprise component of BM	JAR
72	EntMult	Loughran and Wellman	2011	Enterprise Multiple	JFQA
73	EP	Basu	1977	Earnings-to-Price Ratio	JF
74	EquityDuration	Dechow, Sloan and Soliman	2004	Equity Duration	RAS
75	ExchSwitch	Dharan and Ikenberry	1995	Exchange Switch	JF
76	ExclExp	Doyle, Lundholm and Soliman	2003	Excluded Expenses	RAS
77	fgr5yrLag	La Porta	1996	Long-term EPS forecast	JF
78	FirmAgeMom	Zhang	2004	Firm Age - Momentum	JF
79	ForecastDispersion	Diether, Malloy and Scherbina	2002	EPS Forecast Dispersion	JF
80	FR	Franzoni and Marin	2006	Pension Funding Status	JF
81	FRbook	Franzoni and Marin	2006	Pension Funding Status	JF
82	Frontier	Nguyen and Swanson	2009	Efficient frontier index	JFQA
83	G_Binary	Gompers, Ishii and Metrick	2003	Governance Index	QJE
84	GP	Novy-Marx	2013	gross profits / total assets	JFE
85	GrAdExp	Lou	2014	Growth in advertising expenses	RFS
86	grcapx	Anderson and Garcia-Feijoo	2006	Change in capex (two years)	JF

Table D.1. Detailed Descriptions of used Anomalies *(continued)*

No.	Acronym	Authors	Year	Long Description	Journal
87	grcapx1y	Anderson and Garcia-Feijoo	2006	Investment growth (1 year)	AR
88	grcapx3y	Anderson and Garcia-Feijoo	2006	Change in capex (three years)	JF
89	GrGMToGrSales	Abarbanell and Bushee	1998	Gross Margin growth over sales growth	AR
90	GrLTNOA	Fairfield, Whisenant and Yohn	2003	Growth in Long term net operating assets	AR
91	GrSaleToGrInv	Abarbanell and Bushee	1998	Sales growth over inventory growth	AR
92	GrSaleToGrOverhead	Abarbanell and Bushee	1998	Sales growth over overhead growth	AR
93	Herf	Hou and Robinson	2006	Industry concentration (Herfindahl) sales	JF
94	HerfAsset	Hou and Robinson	2006	Industry concentration (Herfindahl) assets	JF
95	HerfBE	Hou and Robinson	2006	Industry concentration (Herfindahl) book	JF
96	High52	George and Hwang	2004	52 week high	JF
97	hire	Bazdresch, Belo and Lin	2014	Employment growth	JPE
98	IdioRisk	Ang et al.	2006	Idiosyncratic risk	JF
99	IdioVol3F	Ang et al.	2006	Idiosyncratic risk (3 factor)	JF
100	IdioVolAHT	Ali, Hwang, and Trombley	2003	Idiosyncratic risk (AHT)	JFE
101	IdioVolCAPM	Ang et al.	2006	Idiosyncratic risk (CAPM)	JF
102	Illiquidity	Amihud	2002	Amihud's illiquidity	JFM
103	IndIPO	Ritter	1991	Initial Public Offerings	JF
104	IndMom	Grinblatt and Moskowitz	1999	Industry Momentum	JFE
105	IndRetBig	Hou	2007	Industry return of big firms	RFS
106	IntanBM	Daniel and Titman	2006	Intangible return using BM	JF
107	IntanCFP	Daniel and Titman	2006	Intangible return using CFtoP	JF
108	IntanEP	Daniel and Titman	2006	Intangible return using EP	JF

Table D.1. Detailed Descriptions of used Anomalies *(continued)*

No.	Acronym	Authors	Year	Long Description	Journal
109	IntanSP	Daniel and Titman	2006	Intangible return using Sale2P	JF
110	IntMom	Novy-Marx	2012	Intermediate Momentum	JFE
111	Investment	Titman, Wei and Xie	2004	Investment to revenue	JFQA
112	InvestPPEInv	Lyandres, Sun and Zhang	2008	change in ppe and inv/assets	RFS
113	InvGrowth	Belo and Lin	2012	Inventory Growth	RFS
114	iomom_cust	Menzly and Ozbas	2010	Customers momentum	JF
115	iomom_supp	Menzly and Ozbas	2010	Suppliers momentum	JF
116	KZ	Lamont, Polk and Saa-Requejo	2001	Kaplan Zingales index	RFS
117	Leverage	Bhandari	1988	Market leverage	JFE
118	MaxRet	Bali, Cakici, and Whitelaw	2010	Maximum return over month	JF
119	MeanRankRevGrowth	Lakonishok, Shleifer and Vishny	1994	Revenue Growth Rank	JF
120	Mom12m	Jegadeesh and Titman	1993	Momentum (12 month)	JF
121	Mom18m13m	De Bondt and Thaler	1985	Momentum-Reversal	JF
122	Mom1m	Jegadeesh	1989	Short term reversal	JF
123	Mom36m	De Bondt and Thaler	1985	Long-run reversal	JF
124	Mom6m	Jegadeesh and Titman	1993	Momentum (6 month)	JF
125	Mom6mJunk	Avramov et al	2007	Junk Stock Momentum	JF
126	MomRev	Chan and Ko	2006	Momentum and LT Reversal	JOIM
127	MomSeas	Heston and Sadka	2008	Return seasonality	JFE
128	MomSeasAlt11to15a	Heston and Sadka	2008	Return seasonality years 11 to 15	JFE
129	MomSeasAlt16to20a	Heston and Sadka	2008	Return seasonality years 16 to 20	JFE
130	MomSeasAlt16to20n	Heston and Sadka	2008	Returns in not-same month years 16 to 20	JFE

Table D.1. Detailed Descriptions of used Anomalies *(continued)*

No.	Acronym	Authors	Year	Long Description	Journal
131	MomSeasAlt1a	Heston and Sadka	2008	Return seasonality last year	JFE
132	MomSeasAlt1n	Heston and Sadka	2008	Returns in not-same month last year	JFE
133	MomSeasAlt2to5n	Heston and Sadka	2008	Returns in not-same years 2 to 5	JFE
134	MomSeasAlt6to10a	Heston and Sadka	2008	Return seasonality years 6 to 10	JFE
135	MomSeasAlt6to10n	Heston and Sadka	2008	Returns in different months years 6 to 10	JFE
136	MomVol	Lee and Swaminathan	2000	Momentum and Volume	JF
137	MS	Mohanram	2005	Mohanram G-score	RAS
138	NetDebtFinance	Bradshaw, Richardson and Sloan	2006	Net debt financing	JAE
139	NetDebtPrice	Penman, Richardson and Tuna	2007	Net debt to price	JAR
140	NetEquityFinance	Bradshaw, Richardson and Sloan	2006	Net equity financing	JAE
141	NetPayoutYield	Boudoukh et al.	2007	Net Payout Yield	JF
142	NOA	Hirshleifer et al.	2004	Net Operating Assets	JAE
143	NumEarnIncrease	Loh and Warachka	2012	Number of consecutive earnings increases	MS
144	OperProf	Fama and French	2006	operating profits / book equity	JFE
145	OperProfRDNoLag	Ball et al.	2016	Cash-based operating profitability	JFE
146	OPLeverage	Novy-Marx	2010	Operating Leverage	ROF
147	OptionVolume1	Johnson and So	2012	Option Volume to Stock Volume	JFE
148	OptionVolume2	Johnson and So	2012	Option Volume relative to recent average	JFE
149	OrderBacklog	Rajgopal, Shevlin and Venkatachalam	2003	Order backlog	RAS
150	OrgCap	Eisfeldt and Papanikolaou	2013	Organizational Capital	JF
151	OrgCapAdj	Eisfeldt and Papanikolaou	2013	Organizational Capital industry adj	JF
152	OScore	Dichev	1998	O Score	JFE

Table D.1. Detailed Descriptions of used Anomalies *(continued)*

No.	Acronym	Authors	Year	Long Description	Journal
153	PatentsRD	Hirschleifer, Hsu and Li	2013	Patents to RD expenses	JFE
154	PayoutYield	Boudoukh et al.	2007	Payout Yield	JF
155	PctAcc	Hafzalla, Lundholm and Van Winkle	2011	Percent Operating Accruals	AR
156	PctTotAcc	Hafzalla, Lundholm and Van Winkle	2011	Percent Total Accruals	AR
157	PredictedFE	Frankel and Lee	1998	Predicted Analyst forecast error	JAE
158	Price	Blume and Husic	1972	Price	JF
159	PriceDelay	Hou and Moskowitz	2005	Price delay coeff	RFS
160	PriceDelayAdj	Hou and Moskowitz	2005	Price delay SE adjusted	RFS
161	PriceDelayRsq	Hou and Moskowitz	2005	Price delay r square	RFS
162	ProbInformedTrading	Easley, Hvidkjaer and O'Hara	2002	Probability of Informed Trading	JF
163	Profitability	Balakrishnan, Bartov and Faurel	2010	Return on assets	JAE
164	PS	Piotroski	2000	Piotroski F-score	AR
165	RD	Chan, Lakonishok and Sougiannis	2001	R&D over market cap	JF
166	RDAbility	Cohen, Diether and Malloy	2013	R&D ability	RFS
167	RDcap	Li	2011	R&D capital-to-assets	RFS
168	RDIPO	Gou, Lev and Shi	2006	IPO and no R&D spending	JBFA
169	RDS	Landsman et al.	2011	Real dirty surplus	AR
170	realestate	Tuzel	2010	Real estate holdings	RFS
171	ResidualMomentum11m	Blitz, Huij and Martens	2011	11 month residual momentum	JEmpFin
172	ResidualMomentum6m	Blitz, Huij and Martens	2011	6 month residual momentum	JEmpFin
173	retConglomerate	Cohen and Lou	2012	Conglomerate return	JFE
174	ReturnSkew	Bali, Engle and Murray	2015	Skewness of daily returns	Book

Table D.1. Detailed Descriptions of used Anomalies *(continued)*

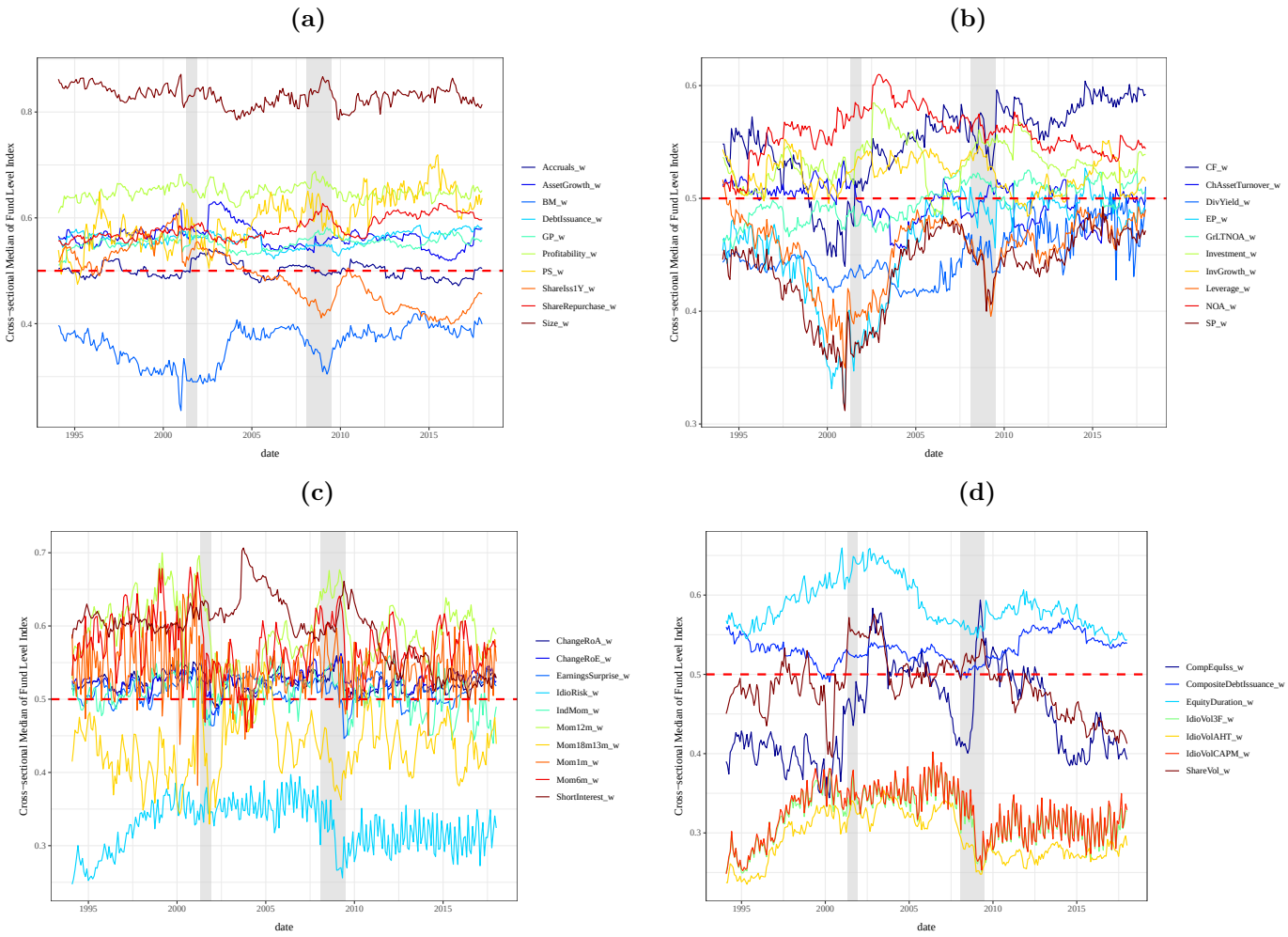
No.	Acronym	Authors	Year	Long Description	Journal
175	ReturnSkew3F	Bali, Engle and Murray	2015	Skewness of daily idiosyncratic returns (3F model)	Book
176	REV6	Chan, Jegadeesh and Lakonishok	1996	Earnings forecast revisions	JF
177	RevenueSurprise	Jegadeesh and Livnat	2006	Revenue Surprise	JFE
178	RIO_BM	Nagel	2005	Inst Own and BM	JF
179	RIO_Disp	Nagel	2005	Inst Own and Forecast Dispersion	JF
180	RIO_IdioRisk	Nagel	2005	Inst Own and Idio Vol	JF
181	RIO_Turnover	Nagel	2005	Inst Own and Turnover	JF
182	roaq	Balakrishnan, Bartov and Faurel	2010	Return on assets incl extraordinary income	JAE
183	secured	Valta	2016	Secured debt	JFQA
184	securedind	Valta	2016	Secured debt indicator	JFQA
185	sfe	Elgers, Lo and Pfeiffer	2001	Earnings Forecast to price	AR
186	ShareIss1Y	Pontiff and Woodgate	2008	Share issuance (1 year)	JF
187	ShareIss5Y	Daniel and Titman	2006	Share issuance (5 year)	JF
188	ShareRepurchase	Ikenberry, Lakonishok and Vermaelen	1995	Share repurchases	JFE
189	ShareVol	Datar, Naik and Radcliffe	1998	Share Volume	JFM
190	ShortInterest	Dechow et al.	2001	Short Interest	JFE
191	sinAlgo	Hong and Kacperczyk	2009	Sin Stock (selection criteria)	JFE
192	Size	Banz	1981	Size	JFE
193	skew1	Xing, Zhang and Zhao	2010	Volatility smirk near the money	JFQA
194	SmileSlope	Yan	2011	Put volatility minus call volatility	JFE
195	SP	Barbee, Mukherji and Raines	1996	Sales-to-price	FAJ
196	Spinoff	Cusatis, Miles and Woolridge	1993	Spinoffs	JFE

Table D.1. Detailed Descriptions of used Anomalies *(continued)*

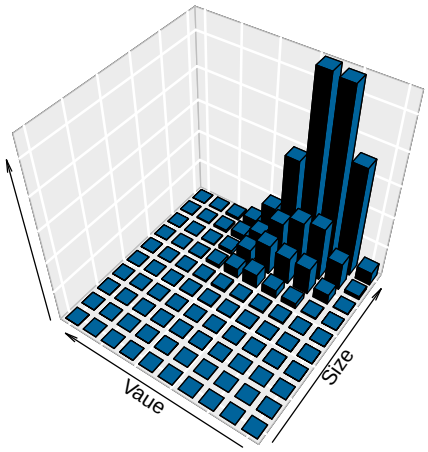
No.	Acronym	Authors	Year	Long Description	Journal
197	std_turn	Chordia, Subrahmanyam and Anshuman	2001	Share turnover volatility	JFE
198	SurpriseRD	Eberhart, Maxwell and Siddique	2004	Unexpected R&D increase	JF
199	tang	Hahn and Lee	2009	Tangibility	JF
200	Tax	Lev and Nissim	2004	Taxable income to income	AR
201	TotalAccruals	Richardson et al.	2005	Total accruals	JAE
202	UpForecast	Barber et al.	2002	Up Forecast	JF
203	VolSD	Chordia, Subrahmanyam and Anshuman	2001	Volume Variance	JFE
204	XFIN	Bradshaw, Richardson and Sloan	2006	Net external financing	JAE
205	zerotrade	Liu	2006	Days with zero trades	JFE
206	zerotradeAlt1	Liu	2006	Days with zero trades	JFE
207	zerotradeAlt12	Liu	2006	Days with zero trades	JFE
208	ZScore	Dichev	1998	Altman Z-Score	JFE

E More Graphical Results for Demonstration

Figure E.1



		Exposure to Value									
		Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Exposure to Size	Q1	0	0	0	0	0	2	1	0	0	1
	Q2	2	0	0	0	0	12	28	0	4	5
	Q3	3	0	0	14	30	159	36	4	20	24
	Q4	7	4	16	83	153	96	105	152	144	38
	Q5	6	47	94	203	256	361	292	142	44	66
	Q6	20	39	454	495	470	712	454	195	64	68
	Q7	25	362	1726	2666	4482	3073	1067	258	159	83
	Q8	238	3228	8635	8771	9682	4756	1202	334	112	72
	Q9	1044	8941	17510	16180	12644	5552	2090	761	346	185
	Q10	4193	35361	55274	56077	28991	10203	2800	774	273	76



Note: In the figure above, we demonstrate distribution of all the mutual funds in our investigated sample over the corresponding fund-level indices constructed as the exposure to the "Size"-related (measured by market equity) and "Value"-related (measured by book-to-market ratio) characteristics of holdings assets. Specifically we equally dissect each fund-level index into 10 quantiles ranging from low to high (denoted by Q_1 to Q_{10}) and then calculate number of funds of different categories characterized by "Size"-related and "Value"-related indices. Each entry of the accompanied $10x10$ matrix demonstrated above collects total number of observed funds in the associated category.

Figure E.2. Cross-sectional distribution of funds on "Size"-related index

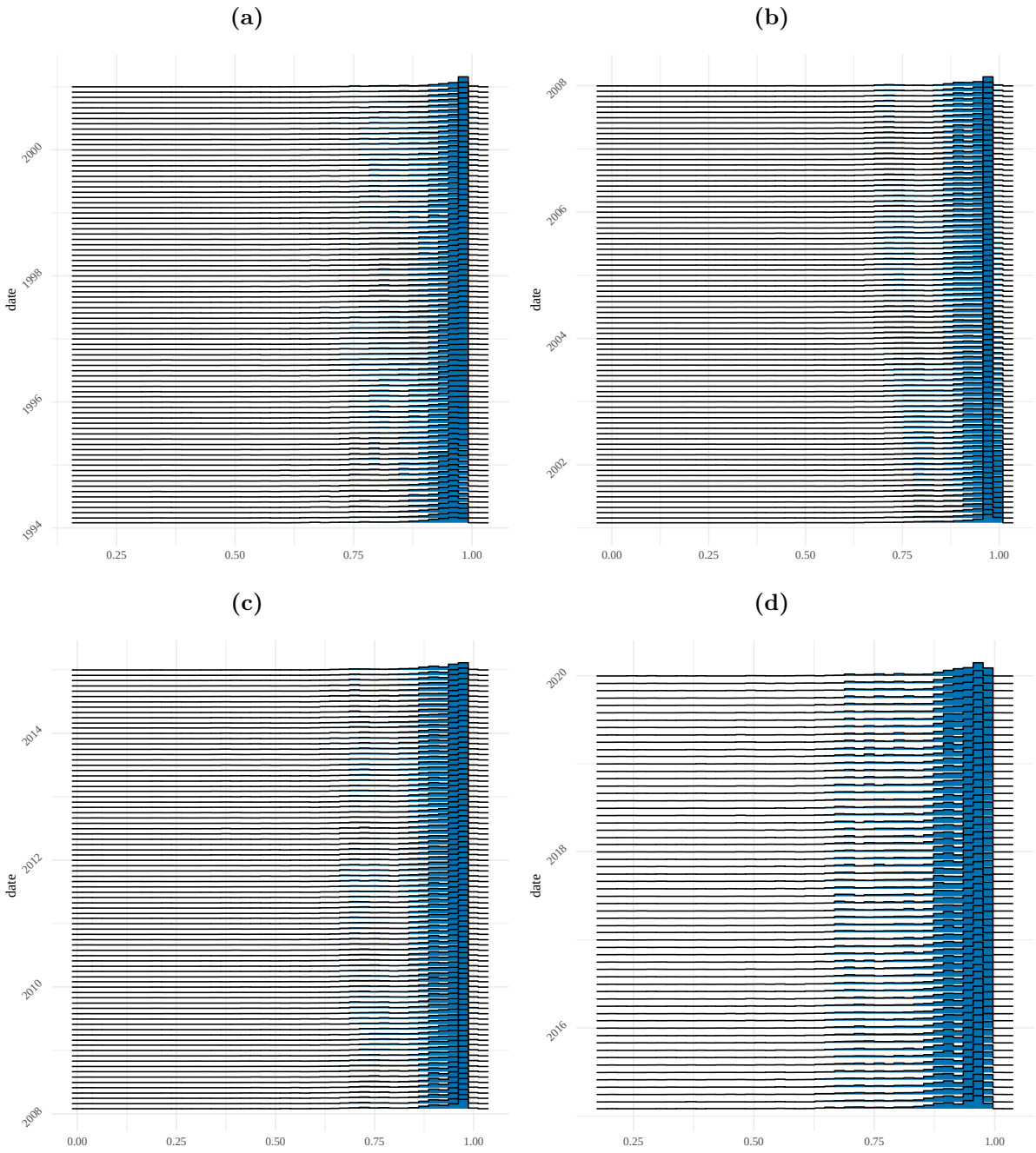


Figure E.3. Cross-sectional distribution of funds on "Value"-related index

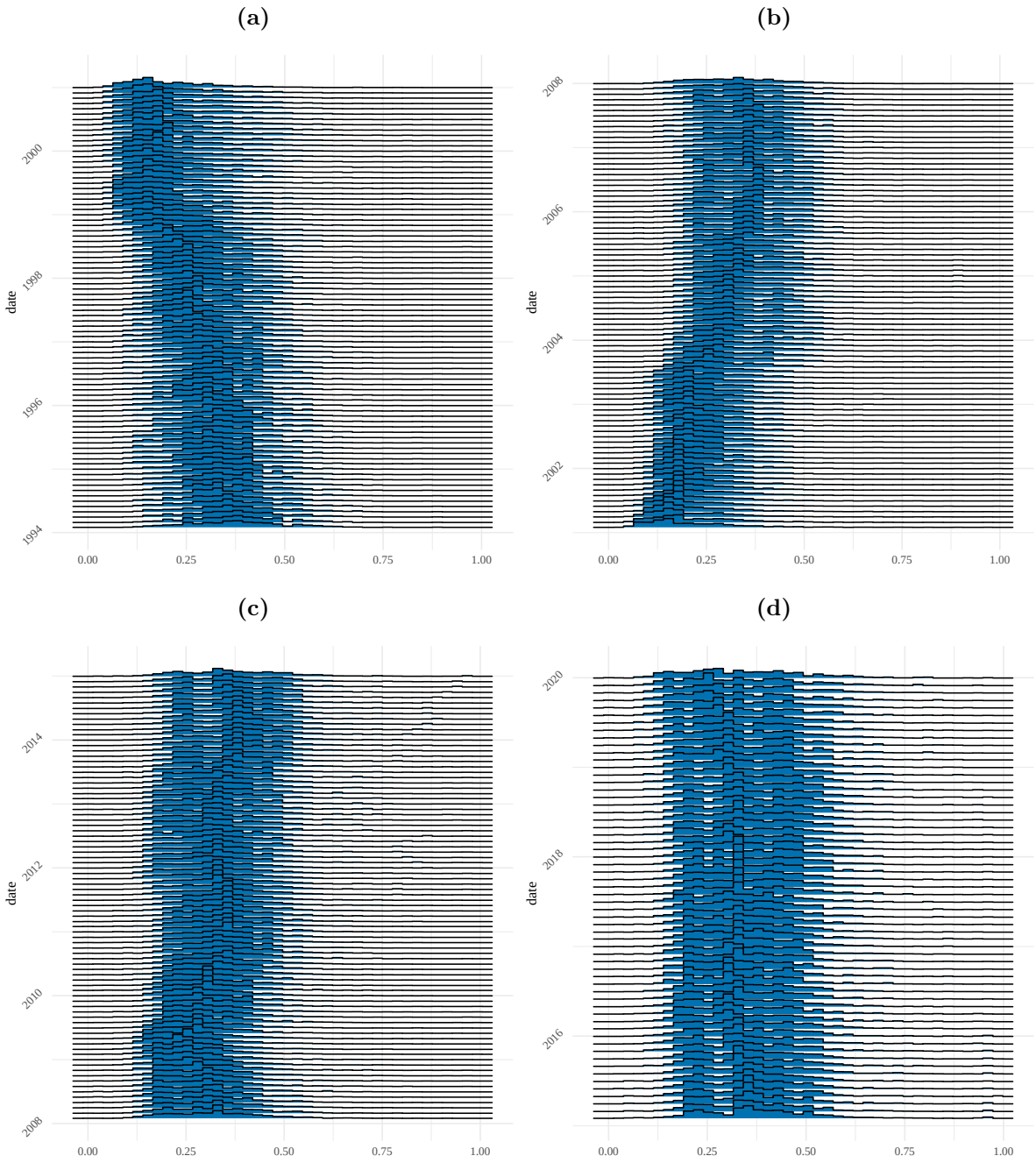
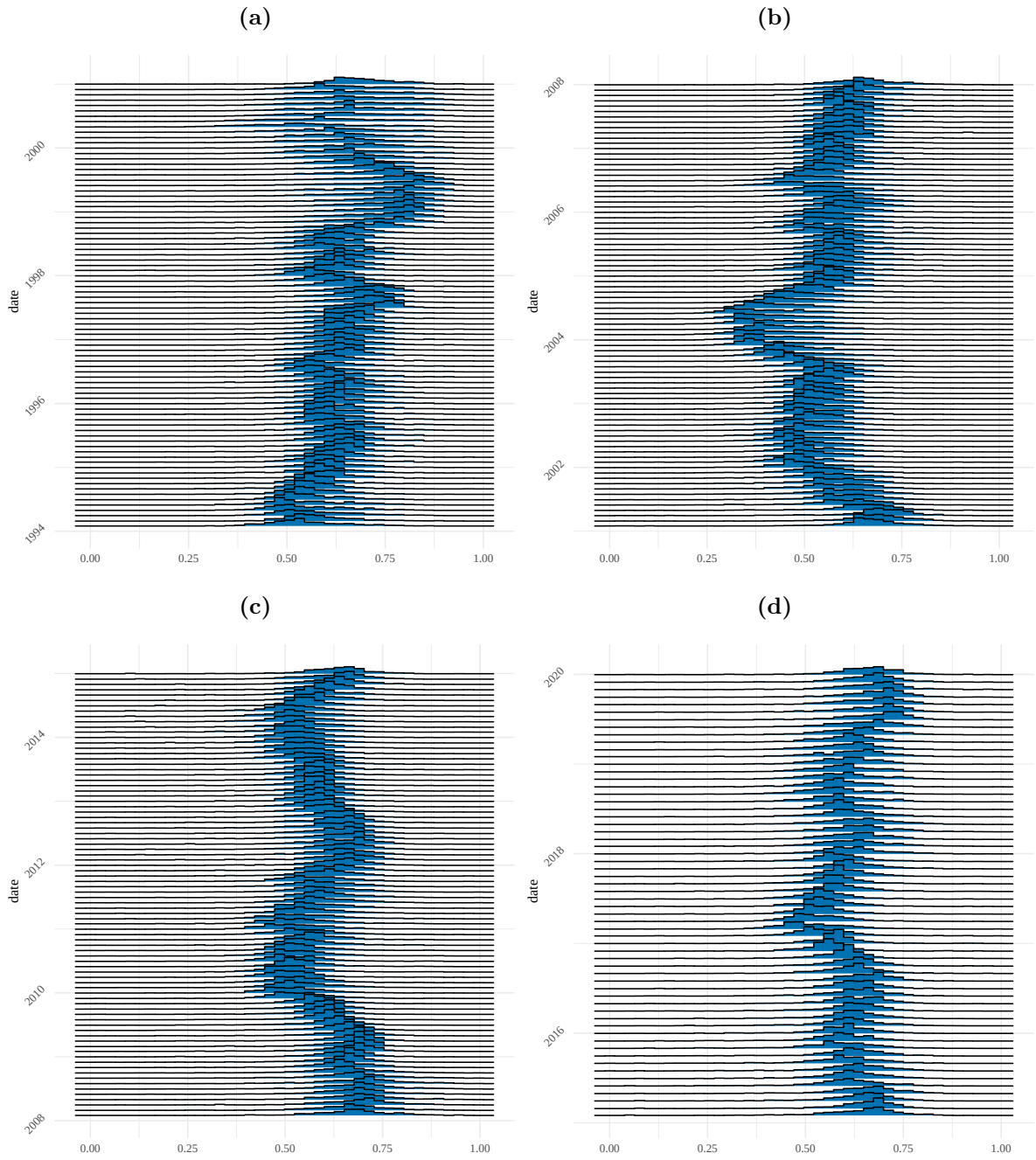


Figure E.4. Cross-sectional distribution of funds on "Momentum"-related index



F One Supplementary Simulation

In this section, we proceed to demonstrate one example illustrating how our proposed ℓ_1/ℓ_q-regularized IPCA works in terms of correctly selecting the corresponding characteristic dimensions that matter to the exposure to factors. The basic procedure simulating data is as following

Step 1. Sample Γ_β with each entry of Γ_β sampled from standard normal distribution, which is a $L \times K$ matrix with L and K specified as $L = 16$ and $K = 3$ respectively.

Step 2. Sample factors as $K \times T$ matrix with each entry sampled from standard uniform distribution.

Step 3. Sample simulated signals (characteristics) as $N \times L \times T$ array with each entry sample from standard uniform distribution. Thus for each sliced matrix with time index fixed at t is $\mathbf{Z}_t$ is a $N \times L$ matrix.

Step 4. Calculate dynamic factor loading as $N \times K$ matrix such that $\mathbf{Z}_t\Gamma_\beta + \texttt{0.01*randn(N,K)}$, where `randn` refers to the standard built-in random number generating function in `MATLAB` for generating random numbers from standard normal distribution.

Step 5. Generate simulated return as specified (4) using simulated $\mathbf{Z}_t$, Γ_β and correspondingly factor loading in the previous steps.

Sample size for our Monte Carlo experiment is specified as $N = 300$ and $T = 360$ and we specif $l = 1, 3, 4, 12, 15, 16$ as the indices for rows of Γ_β that are equipped with non-zero ℓ_2-norm. The suggested indices based on the non-zero ℓ_2-norm of row vectors of estimated $\hat{\Gamma}_\beta$ are $l = 1, 3, 4, 12, 15, 16$.

Γ_β

1	0.465	0.468	0.463
2	0.000	0.000	0.000
3	-0.694	0.115	0.408
4	-0.455	-0.247	0.100
5	0.000	0.000	0.000
6	0.000	0.000	0.000
7	0.000	0.000	0.000
8	0.000	0.000	0.000
9	0.000	0.000	0.000
10	0.000	0.000	0.000
11	0.000	0.000	0.000
12	0.302	-0.811	0.221
13	0.000	0.000	0.000
14	0.000	0.000	0.000
15	0.041	-0.211	0.477
16	0.047	0.073	0.577

$\hat{\Gamma}_\beta$

1	0.821	-0.088	0.030
2	0.000	0.000	0.000
3	-0.071	-0.343	-0.627
4	-0.320	-0.449	-0.357
5	0.000	0.000	0.000
6	0.000	0.000	0.000
7	0.000	0.000	0.000
8	0.000	0.000	0.000
9	0.000	0.000	0.000
10	0.000	0.000	0.000
11	0.000	0.000	0.000
12	-0.178	-0.646	0.673
13	0.000	0.000	0.000
14	0.000	0.000	0.000
15	0.160	-0.304	0.013
16	0.402	-0.404	-0.162

	Bh				
	1		2		
A	B	C	D	E	F
1	0.356	0.648	0.169	0.441	-1.421
2	-1.6222	-0.336	0.764	-0.500	0.431
3	0.883	0.130	0.851	1.495	4.053
4	-0.790	-1.334	0.286	-0.119	0.398
5	-0.779	-1.708	-0.129	-0.543	-0.682